Coling 2010

23rd International Conference on Computational Linguistics

Proceedings of the

2nd Workshop on The People's Web Meets NLP: Collaboratively Constructed Semantic Resources

Produced by
Chinese Information Processing Society of China
No.4 Zhong Guan Cun Nan Si Jie, Hai Dian District
Beijing, 100084
China

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
`acl@aclweb.org`

Introduction

This volume contains papers accepted for presentation at the 2nd Workshop on Collaboratively Constructed Semantic Resources that took place on August 28, 2010, as part of the Coling 2010 conference in Beijing. Being the second workshop on this topic, we were able to build on the success of the previous workshop on this topic held as part of ACL-IJCNLP 2009.

In many works, collaboratively constructed semantic resources have been used to overcome the knowledge acquisition bottleneck and coverage problems pertinent to conventional lexical semantic resources. The greatest popularity in this respect can so far certainly be attributed to Wikipedia. However, other resources, such as folksonomies or the multilingual collaboratively constructed dictionary Wiktionary, have also shown great potential. Thus, the scope of the workshop deliberately includes any collaboratively constructed resource, not only Wikipedia.

Effective deployment of such resources to enhance Natural Language Processing introduces a pressing need to address a set of fundamental challenges, e.g. the interoperability with existing resources, or the quality of the extracted lexical semantic knowledge. Interoperability between resources is crucial as no single resource provides perfect coverage. The quality of collaboratively constructed semantic resources is a fundamental issue, as they lack editorial control and entries are often incomplete. Thus, techniques for link prediction or information extraction have been proposed to guide the ”crowds” while constructing resources of better quality.

We issued calls for both long and short papers. Seven long papers and one short paper were accepted for presentation, based on the careful reviews of our program committee. We would like to thank the program committee members for their thoughtful, high quality, and elaborate reviews, especially considering the tight schedule for reviewing. The call for papers attracted submissions on a wide range of topics showing that collaboratively constructed semantic resources are of growing interest in different fields of Natural Language Processing.

The workshop aimed at bringing together researchers from different worlds, for example those using collaboratively constructed resources as sources of lexical semantic information for Natural Language Processing purposes such as information retrieval, named entity recognition, or keyword extraction, and those using Natural Language Processing techniques to improve the resources or extract and analyze different types of lexical semantic information from them. Looking at the final proceedings, we can safely say that this goal has been achieved.

Iryna Gurevych and Torsten Zesch

Organizers:

Iryna Gurevych, UKP Lab, Technische Universität Darmstadt
Torsten Zesch, UKP Lab, Technische Universität Darmstadt

Program Committee:

Andras Csomai, Google Inc.
Anette Frank, Heidelberg University
Benno Stein, Bauhaus University Weimar
Bernardo Magnini, ITC-irst Trento
Christiane Fellbaum, Princeton University
Dan Moldovan, University of Texas at Dallas
Delphine Bernhard, LIMSI-CNRS, Orsay
Diana McCarthy, Lexical Computing Ltd
Elke Teich, Technische Universität Darmstadt
Emily Pitler, University of Pennsylvania
Eneko Agirre, University of the Basque Country
Erhard Hinrichs, Eberhard Karls Universitt Tübingen
Ernesto De Luca, Technische Universität Berlin
Florian Laws, University of Stuttgart
Gerard de Melo, MPI Saarbrücken
German Rigau, University of the Basque Country
Graeme Hirst, University of Toronto
Günter Neumman, DFKI Saarbrücken
Gy¨rgy Szarvas, Technische Universität Darmstadt
Hans-Peter Zorn, European Media Lab, Heidelberg
Jos Iria, University of Sheffield
Laurent Romary, LORIA, Nancy
Magnus Sahlgren, Swedish Institute of Computer Science
Manfred Stede, Potsdam University
Omar Alonso, Microsoft
Pablo Castells, Universidad Autnonoma de Madrid
Paul Buitelaar, DERI, Galway
Philipp Cimiano, Delft University of Technology
Razvan Bunescu, University of Texas at Austin
Rene Witte, Concordia University Montral
Roxana Girju, University of Illinois at Urbana-Champaign
Saif Mohammad, University of Maryland
Samer Hassan, University of North Texas
Sren Auer, Leipzig University
Tonio Wandmacher, CEA, Paris

Invited Speaker:

Tat-Seng Chua, National University of Singapore

Title: Extracting Knowledge from Community Question-Answering Sites

Abstract: Community question-answering (QA) services, like Yahoo! Answers, contain a huge amount of information in the form of QA pairs accumulated over many years. The information covers a wide variety of topics on questions of great interests to and frequently asked by the users. To make this huge amount of information accessible by general users, research has been carried out to help users find similar questions with readily available answers. However, a better approach is to organize all relevant QA pairs around a given topic into a knowledge structure to help users better understand the overall topic. To accomplish this, our research leverages on appropriate topic prototype hierarchy automatically acquired from the Web or Wikipedia to guide the organization of the un-structured user-generated-contents in community QA sites. More specifically, we propose a prototype-hierarchy based clustering algorithm that utilizes the category structure information, article contents of Wikipedia, as well as distribution of relevant QA pairs around the topic based on a multi-criterion optimization function. This talk discusses our research to transform unstructured community QA resources into knowledge structure.

Short Bio: Chua Tat-Seng the KITHC Chair Professor at the School of Computing, National University of Singapore (NUS). He was the Acting and Founding Dean of the School of Computing during 1998-2000. He joined NUS in 1983, and spent three years as a research staff member at the Institute of Systems Science (now I2R) in the late 1980s. Dr Chua's main research interest is in multimedia information retrieval, in particular, on the analysis, retrieval and question-answering (QA) of text and image/video information. He is currently working on several multi-million-dollar projects: interactive media search, local contextual search, and real-time live media search. His group participates regularly in TREC-QA and TRECVID video retrieval evaluations. Dr Chua has organized and served as program committee member of numerous international conferences in the areas of computer graphics, multimedia and text processing. He is the conference co-chair of ACM Multimedia 2005, CIVR (Conference on Image and Video Retrieval) 2005, and ACM SIGIR 2008. He serves in the editorial boards of:ACM Transactions of Information Systems (ACM), Foundation and Trends in Information Retrieval (NOW), The Visual Computer (Springer Verlag), and Multimedia Tools and Applications (Kluwer). He is the member of steering committee of CIVR, Computer Graphics International, and Multimedia Modeling conference series; and as member of International Review Panels of two large-scale research projects in Europe.

Table of Contents

Conference Program

Saturday, August 28, 2010

9:15–9:30 Opening Remarks

9:30–10:00 *Constructing Large-Scale Person Ontology from Wikipedia*
Yumi Shibaki, Masaaki Nagata and Kazuhide Yamamoto

10:00–10:30 *Using the Wikipedia Link Structure to Correct the Wikipedia Link Structure*
Benjamin Mark Pateman and Colin Johnson

10:30–11:00 Coffee Break

11:00–11:30 *Extending English ACE 2005 Corpus Annotation with Ground-truth Links to Wikipedia*
Luisa Bentivogli, Pamela Forner, Claudio Giuliano, Alessandro Marchetti, Emanuele Pianta and Kateryna Tymoshenko

11:30–12:00 *Expanding textual entailment corpora fromWikipedia using co-training*
Fabio Massimo Zanzotto and Marco Pennacchiotti

12:00–12:30 *Pruning Non-Informative Text Through Non-Expert Annotations to Improve Aspect-Level Sentiment Classification*
Ji Fang, Bob Price and Lotti Price

12:30–14:00 Lunch Break

14:00–15:00 Invited Talk by Tat-Seng Chua, National University of Singapore

15:00–15:30 *Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure*
Stephan Gouws, G-J van Rooyen and Herman A. Engelbrecht

15:30–16:00 Coffee Break

16:00–16:30 *Identifying and Ranking Topic Clusters in the Blogosphere*
M. Atif Qureshi, Arjumand Younus, Muhammad Saeed, Nasir Touheed, Emanuele Pianta and Kateryna Tymoshenko

16:30–16:50 *Helping Volunteer Translators, Fostering Language Resources*
Masao Utiyama, Takeshi Abekawa, Eiichiro Sumita and Kyo Kageura

Saturday, August 28, 2010 (continued)

16:50–17:30 Discussion

Constructing Large-Scale Person Ontology from Wikipedia

Yumi Shibaki
Nagaoka University of Technology
shibaki@jnlp.org

Masaaki Nagata
NTT Communication Science Laboratories
nagata.masaaki@labs.ntt.co.jp

Kazuhide Yamamoto
Nagaoka University of Technology
yamamoto@jnlp.org

Abstract

This paper presents a method for constructing a large-scale Person Ontology with category hierarchy from Wikipedia. We first extract Wikipedia category labels which represent person (hereafter, Wikipedia Person Category, WPC) by using a machine learning classifier. We then construct a WPC hierarchy by detecting *is-a* relations in the Wikipedia category network. We then extract the titles of Wikipedia articles which represent person (hereafter, Wikipedia person instance, WPI). Experiments show that the accuracy of WPC extraction is 99.3% precision and 98.4% recall, while that of WPI extraction is 98.2% and 98.6%, respectively. The accuracies are significantly higher than the previous methods.

1 Introduction

In recent years, we have become increasingly aware of the need for, up-to-date knowledge bases offering broad coverage in order to implement practical semantic inference engines for advanced applications such as question answering, summarization and textual entailment recognition. General ontologies, such as WordNet (Fellbaum et al., 1998), and *Nihongo Goi-Taikei* (Ikehara et al., 1997), contain general knowledge of wide range of fields. However, it is difficult to instantly add new knowledge, particularly proper nouns, to these general ontologies. Therefore, Wikipedia has come to be used as a useful corpus for knowledge extraction because it is a free and large-scale online encyclopedia that continues to be actively developed. For example, in DBpedia (Bizer et al. 2009), RDF triples are extracted from the Infobox templates within Wikipedia articles. In YAGO (Suchanek et al. 2007), an appropriate WordNet synset (most likely category) is assigned to a Wikipedia category as a super-category, and Wikipedia articles are extracted as instances of the category.

As a first step to make use of proper noun and related up-to-date information in Wikipedia, we focus on person names and the articles and categories related to them because it contains a large number of articles and categories that indicate person, and because large-scale person ontology is useful for applications such as person search and named entity recognition. Examples of a person article are personal name and occupational title such as “Ichiro” and “Financial planner,” while an example of a person category is occupational title such as “Sportspeople.”

The goal of this study is to construct a large-scale and comprehensive person ontology by extracting person categories and *is-a* relations[1] among them. We first apply a classifier based on machine learning to all Wikipedia categories to extract categories that represent person. If both of the linked Wikipedia categories are person categories, the category link is labeled as an *is-a* relation. We then use a heuristic-based rule to extract the title of articles that represent person as person instance from the person categories.

In the following sections, we first describe the language resources and the previous works. We then introduce our method for constructing the person ontology and report our experimental results.

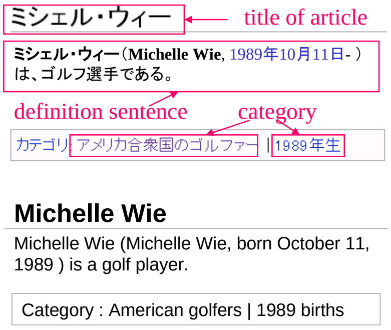
[1] “*is-a* relation” is defined as a relation between A and B when “B is a (kind of) A.”

Beijing, August 2010
Proceedings of the 2nd Workshop on “Collaboratively Constructed Semantic Resources”, Coling 2010, pages 1–9,

2 Language Resources

2.1 Japanese Wikipedia

Wikipedia is a free, multilingual, on-line encyclopedia that is being actively developed by a large number of volunteers. Wikipedia has articles and categories. The data is open to the public as XML files[2]. Figure 1 shows an example of an article. An article page has a title, body, and categories. In most articles, the first sentence of the body is the definition sentence of the title. Although the Wikipedia category system is organized in a hierarchal manner, it is a thematic classification, not a taxonomy. The relation between category and subcategory and that between a category and articles listed on it are not necessarily an *is-a* relation. A category could have two or more super categories and the category network could have loops.



ミシェル・ウィー

ミシェル・ウィー（**Michelle Wie**, 1989年10月11日- ）は、ゴルフ選手である。

カテゴリ: アメリカ合衆国のゴルファー | 1989年生

Michelle Wie

Michelle Wie (Michelle Wie, born October 11, 1989) is a golf player.

Category : American golfers | 1989 births

Figure 1: Example of title, body (definition sentence), and categories for article page in Japanese Wikipedia (top) and its translation (bottom)

2.2 Nihongo Goi-Taikei

To construct the ontology, we first apply a machine learning based classifier to determine if a category label indicates a person or not. A Wikipedia category label is often a common compound noun or a noun phrase, and the head word of a Japanese compound noun and noun phrase is usually the last word. We assume the semantic category of the last word is an important feature for classification.

Nihongo Goi-Taikei (hereafter, *Goi-Taikei*) is one of the largest and best known Japanese thesauri. *Goi-Taikei* contains different semantic category hierarchies for common nouns, proper nouns, and verbs. In this work, we use only the common noun category (Figure 2). It consists of approximately 100,000 Japanese words (hereafter, instance) and the meanings of each word are described by using about 2,700 hierarchical semantic categories. Words (Instances) with multiple meanings (ambiguous words) are assigned multiple categories in *Goi-Taikei*. For example, the transliterated Japanese word (instance) *raita* (ライター) has two meanings of "writer" and "lighter," and so belongs to two categories, "353:author[3]" and "915:household."

Japanese WordNet (approximately 90,000 entries as of May 2010), which has recently been released to the public (Bonds et al., 2008), could be an alternative to *Goi-Taikei* as a large-scale Japanese thesaurus. We used *Goi-Taikei* in this work because Japanese WordNet was translated from English WordNet and it is not known whether it covers the concepts unique to Japanese.

3 Previous Works

3.1 Ponzetto's method and Sakurai's method

Ponzetto et al. (2007) presented a set of lightweight heuristics such as head matching and modifier matching for distinguishing *is-a* links from *not-is-a* links in the Wikipedia category network. The main heuristic, "Syntax-based methods" is based on head matching, in which a category link is labeled as *is-a* relation if the two categories share the same head lemma, such as CAPITALS IN ASIA and CAPITALS. Sakurai et al. (2008) presented a method equivalent to head matching for Japanese Wikipedia. As Japanese is a head final language, they introduced the heuristic called *suffix matching*; it labels a category link as a *is-a* relation if one category is the suffix of the other category, such as 日本の空港(airports in Japan) and 空港(airports). In the proposed method herein, if a Wikipedia category and its parent category are both person categories, the category link is labeled as *is-a* relation. Therefore, *is-a* relations, which cannot be extracted by Ponzetto's or Sakurai's method, can be extracted.

[2] http://download.wikimedia.org/jawiki

[3] The *Goi-Taikei* category is prefixed with ID number.

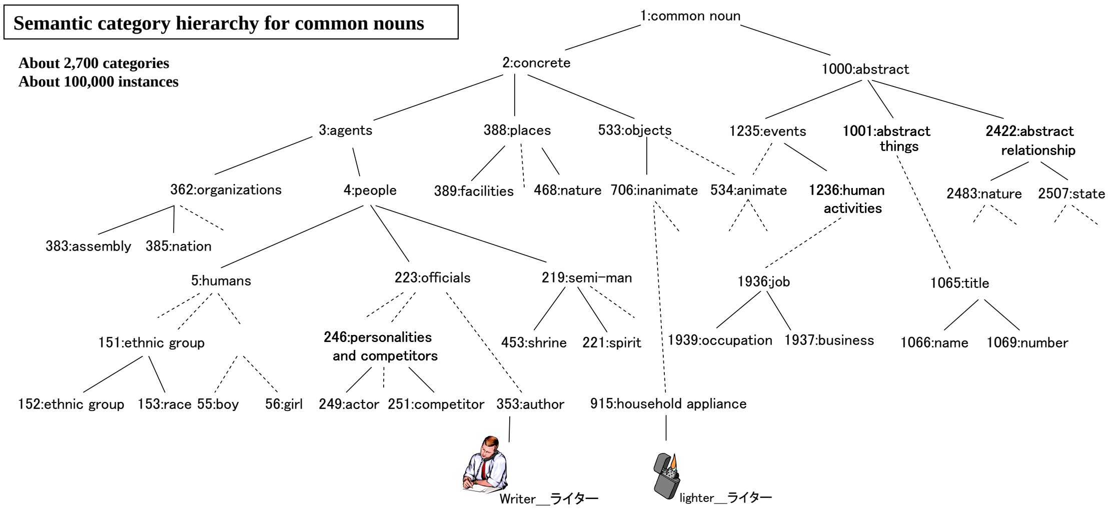


Figure 2: Part of a category hierarchy for common nouns in *Nihongo Goi-Taikei*

3.2 Kobayashi's method

Kobayashi et al. (2008) presented a technique to make a Japanese ontology equivalent to YAGO; it assigns *Goi-Taikei* categories to Japanese Wikipedia categories. These two methods and our method are similar in that a Wikipedia category and the title of an article are regarded as a category and an instance, respectively. Kobayashi et al. automatically extract hypernyms from the definition sentence of each article in advance (referred to hereafter as "*D-hypernym*.") They apply language-dependent lexico-syntactic patterns to the definition sentence to extract the *D-hypernym*. Here are some examples.

は、[hypernym]の一つである <EOS>
one of [hypernym]

は、[hypernym]である<EOS>
is a [hypernym]

[hypernym] <EOS>
is a [hypernym] …

where <EOS> refer to the beginning of a sentence

For example, from the article in Figure 1, the words "ゴルフ選手 (golf player)" is extracted as the *D-hypernym* of the article "ミシェル・ウィー (Michelle Wie)."

Figure 3 outlines the Kobayashi's method. First, for a Wikipedia category, if its last word matches an instance of *Goi-Taikei* category, all such *Goi-Taikei* categories are extracted as a candidate of the Wikipedia category's super-class. If the last word of the *D-hypernym* of the Wikipedia article listed on the Wikipedia category matches an instance of the *Goi-Taikei* category, the *Goi-Taikei* category is extracted as the super-class of the Wikipedia category and its instances (Wikipedia articles) (Figure 3). Although the Kobayashi's method is a general one, it can be used to construct person ontology if the super-class candidates are restricted to those *Goi-Taikei* categories which represent person.

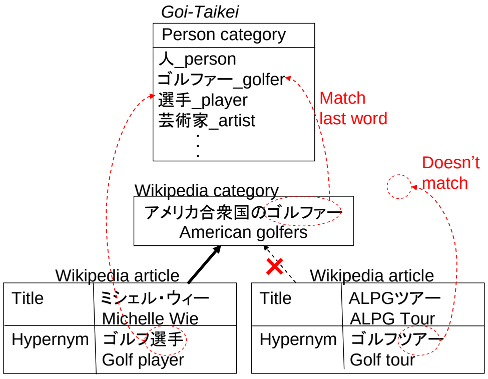


Figure 3: The outline of Kobayashi's method

3.3 Yamashita's method

Yamashita made an open source software which extracts personal names from Japanese Wikipedia[4]. He extracted the titles of articles listed on the categories ○年生(○ births) (e.g., 2000 births). As these categories are used to sort the names of people, horses, and dogs by born year, he used a simple pattern matching

[4] http://coderepos.org/share/browser/lang/perl/misc/wikipejago

rules to exclude horses and dogs. In the experiment in Section 5, we implemented his method by using not only "年生 (births)" but also "年没 (deaths)" and "世紀没 (th-century deaths)," "年代没 (s deaths)," "年代生 (s births)," and "世紀生 (th births)" to extract personal names. As far as we know, it is the only publicly available software to extract a large number of person names from the Japanese Wikipedia. For the comparison with our method, it should be noted that his method cannot extract person categories.

4 Ontology Building Method

4.1 Construction of Wikipedia person category hierarchy (WPC)

We extract the WPC by using a machine learning classifier. If a Wikipedia category and its parent category are both person categories, the category link is labeled as an *is-a* relation. This means that all *is-a* relations in our person ontology are extracted from the original Wikipedia category hierarchy using only a category classifier. This is because we investigated 1,000 randomly sampled links between person categories and found 98.7% of them were *is-a* relations. Figure 4 shows an example of the Wikipedia category hierarchy and the constructed WPC hierarchy.

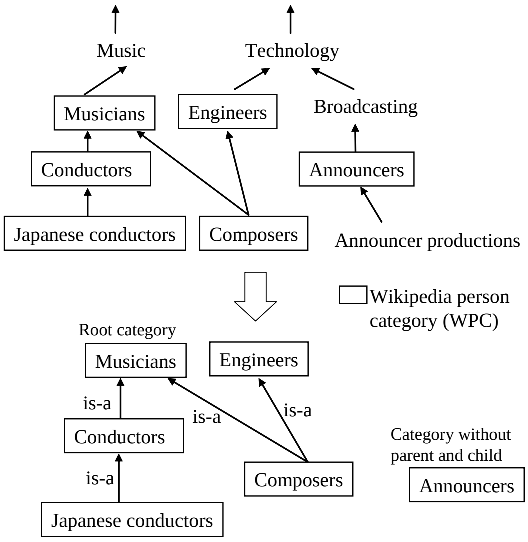


Figure 4: Example of Wikipedia category hierarchy (top) and constructed Wikipedia person category hierarchy (bottom)

We detect whether the Wikipedia category label represents a person by using Support Vector Machine (SVM). The semantic category of the words in the Wikipedia category label and those in the neighboring categories are used for the features. We use the following three aspects of the texts that exist around the target category for creating the features:

1. Structural relation between the target category and the text in Wikipedia. (6 kinds)
2. Span of the text. (2 kinds)
3. Semantic category of the text derived from *Goi-Taikei*. (4 kinds)

We examined 48 features by combining the above three aspects (6*2*4).

The following are the six structural relations in Wikipedia between the target category and the text information:

Structural relation

A. The target Wikipedia category label.

B. All parent category labels of the target category.

C. All child category labels of the target category.

D. All sibling category labels of the target category.

E. All *D-hypernym*[5] from each article listed on the target category.

F. All *D-hypernyms* extracted from the articles with the same name as the target category.

As for F, for example, when the article ベーシスト(bassist) is listed on the category: ベーシスト (bassist), we regard the *D-hypernym* of the article as the hypernym of the category.

As most category labels and *D-hypernyms* are common nouns, they are likely to match instances in *Goi-Taikei* which lists possible semantic categories of words.

[5] As for *D-hypernym* extraction patterns, we used almost the same patterns described in previous works on Japanese sources such as (Kobayashi et al. 2008; Sumida et al., 2008), which are basically equivalent to the works on English sources such as (Hearst, 1992).

After the texts located at various structural relations A-F are collected, they are matched to the instances of *Goi-Taikei* in two different spans:

Span of the text

Ⅰ. All character strings of the text

Ⅱ. The last word of the text

For the span Ⅱ, the text is segmented into words using a Japanese morphological analyzer. The last word is used because the last word usually represents the meaning of the entire noun phrase (semantic head word) in Japanese.

In the proposed method, hierarchical semantic categories of *Goi-Taikei* are divided into two categories; "*Goi-Taikei* person categories" and other categories. *Goi-Taikei* person category is defined as those categories that represent person, that is, all categories under "5:humans" and "223:officials," and "1939: occupation" and "1066:name" in *Goi-Taikei* hierarchy as shown in Figure 1.

For each structural relation A-F and span Ⅰ and Ⅱ, we calculate four relative frequencies a-d, which represents the manner in which the span of texts match the instance of *Goi-Taikei* person category. It basically indicates the degree to which the span of text is likely to mean a person.

Semantic type

a. The span of text matches only instances of *Goi-Taikei* person categories.

b. The span of text matches only instances of categories other than *Goi-Taikei* person categories.

c. The span of text matches both instances of *Goi-Taikei* person categories and those of other categories.

d. The span of text does not match any instances of *Goi-Taikei*.

For example, when the target category is "音楽家" (musicians) in Figure 5 and the feature in question is B-Ⅱ (the last word of its parent categories), the word "家" (whose senses are family and house) falls into semantic type c, and the word "音楽" (music) falls into semantic type b. Therefore, the frequency of semantic types a, b, c, d are 0, 1, 1, 0, respectively, in the features related to B-Ⅱ, and the relative frequencies used for the feature value related B-Ⅱ are 0, 0.5, 0.5, 0, respectively. In this way, we use 48 relative frequencies calculated from the combinations of structural relation A-F, span Ⅰ and Ⅱ, and semantic type a-d, as the feature vector for the SVM.

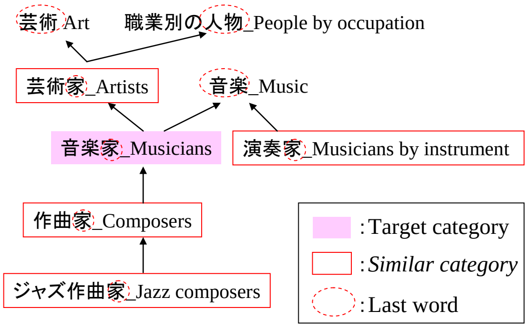


Figure 5: Example of Wikipedia category hierarchy when the target category is "音楽家"

4.2 *Similar category*

In Wikipedia, there are categories that do not have articles and those with few neighboring categories. Here, we define the neighboring categories for a category as those categories that can be reached through a few links from the category. In these cases, there is a possibility that there is not enough text information from which features (mainly semantic category of words) can be extracted, which could degrade the accuracy.

The proposed method overcomes this problem by detecting categories similar to the target category (the category in question) from its neighboring categories for extracting sufficient features to perform classification. Here, "*similar category*" is defined as parent, child, and sibling categories whose last word matches the last word of the target category. This is because there is a high possibility that the *similar categories* and the target category have similar meaning if they share the same last word in the category labels. If the parent (child) category is determined as a *similar category*, its parent (child) category is also determined as a *similar category* if the last word is the same. The procedure is repeated as long as they share the same last word.

Figure 5 shows an example of *similar categories* when the target category is "Musicians." In this case, features extracted from A-F of *similar categories* are added to features extracted using A-F of the target category, “Musicians.” For example, *similar category* “Artists” has “Art” and “People by occupation” as B (parent categories of the target category) in Figure 5, therefore “Art” and “People by occupation” are added to B of “Musicians.”

4.3 Extracting Wikipedia person instance (WPI)

The proposed method extracts, as WPIs the titles of articles listed as WPCs that meet the following four requirements.

1. The last word of the *D-hypernym* of the title of the Wikipedia article matches an instance of *Goi-Taikei* person category.
2. The last word of the title of Wikipedia article matches an instance of Goi-Taike person category.
3. At least one of the Wikipedia categories assigned to the Wikipedia article matches the following patterns:

   (年没|世紀没|年代没|年生|世紀生|年代生)<EOS>
   (deaths | th-century deaths | ’s deaths | births | th-births | ’s births) <EOS>

   These categories are used to sort a large number of person names by year.

4. Wikipedia categories assigned to the Wikipedia article satisfy the following condition:

   $$\frac{\text{Number of extracted WPCs in Section 4.1}}{\text{All number of Wikipedia categories}} > 0.5$$

   This condition is based on the observation that the more WPCs a Wikipedia article is assigned to, the more it is likely to be a WPI. We set the threshold 0.5 from the results of a preliminary experiment.

5 Experiments

5.1 Experimental setup

We used the XML file of the Japanese Wikipedia as of July 24, 2008. We removed irrelevant pages by using keywords (e.g., “image:,” “Help:”) in advance. This cleaning yielded 477,094 Wikipedia articles and 39,782 Wikipedia categories. We manually annotated each category to indicate whether it represents person (positive) or not (negative). For ambiguous cases, we used the following criteria:

＊Personal name by itself (e.g., Michael Jackson) is not regarded as WPC because usually it does not have instances. (Note: personal name as article title is regarded as WPI.)

＊Occupational title (e.g., Lawyers) is regarded as WPC because it represents a person.

＊Family (e.g., Brandenburg family) and Ethnic group (e.g., Sioux) are regarded as WPC.

＊Group name (e.g., The Beatles) is not regarded as WPC.

In order to develop a person category classifier, we randomly selected 2,000 Wikipedia categories (positive:435, negative:1,565) from all categories for training[6]. We used the remaining 37,767 categories for evaluation. To evaluate WPI extraction accuracy, we used Wikipedia articles not listed on the Wikipedia categories used for training. 417,476 Wikipedia articles were used in the evaluation.

To evaluate our method, we used TinySVM-0.09[7] with a linear kernel for classification, and the Japanese morphological analyzer JUMAN-6.0[8] for word segmentation. The comparison methods are Kobayashi’s method and Yamashita’s method under the same conditions as our method.

5.2 Experimental results

Table 1 shows the WPCs extraction accuracy. Precision and recall of proposed method are 6.5 points and 14.8 points better than those of Kobayashi's method, respectively.

	Precision	Recall	F-measure
Kobayashi’s method	92.8% (6727/7247)	83.6% (6727/8050)	88.0%
Proposed method	**99.3%** (7922/7979)	**98.4%** (7922/8050)	**98.8%**

Table 1: The Wikipedia person categories (WPCs) extraction accuracy

[6] We confirmed that the accuracy will level off about 2, 000 training data by experiment. Details will be described in Section 6.

[7] http://chasen.org/~taku/software/TinySVM/

[8] http://www-lab25.kuee.kyoto-u.ac.jp/nl-resource/juman.html

To confirm our assumption on the links between WPCs, we randomly selected 1,000 pairs of linked categories from extracted WPCs, and manually investigated whether both represented person and were linked by *is-a* relation. We found that precision of these pairs was 98.3%.

Errors occurred when the category link between person categories in the Wikipedia category network was not an *is-a* relation, such as 千葉氏(*Chiba* clan) – 大須賀氏(*Ohsuga* clan). However, this case is infrequent, because 98.7% of the links between person categories did exhibit an *is-a* relation (as described in Section 4.1).

Table 2 shows the WPIs extraction accuracy. We randomly selected 1,000 Wikipedia articles from all categories in Wikipedia, and manually created evaluation data (positive:281, negative:719). The recall of the proposed method was 98.6%, 21.0 points higher than that of Yamashita’s method. Our method topped the F-measure of Kobayashi’s method by 3.4 points. Among 118,552 extracted as WPIs by our method, 116,418 articles were expected be correct. In our method, errors occurred when WPI was not listed on any WPCs. However, this case is very rare. Person instances are almost always assigned to at least one WPC. Thus, we can achieve high coverage for WPIs even if we focus only on WPCs. We randomly selected 1,000 articles from all articles and obtained 277 person instances by a manual evaluation. Furthermore, we investigated the 277 person instances, and found that only two instances were not classified into any WPCs (0.7%).

	Precision	Recall	F-measure
Yamashita's method	**100.0%** (218/218)	77.6% (218/281)	87.4%
Kobayashi's method	96% (264/275)	94.0% (264/281)	95.0%
Proposed method	98.2% (277/282)	**98.6%** (277/281)	**98.4%**

Table 2: The Wikipedia person instance (WPIs) extraction accuracy

Table 3 shows the extracted WPC-WPI pairs (e.g., American golfers-Michelle Wie, Artists-Meritorious Artist) extraction accuracy. We randomly selected 1,000 pairs of Wikipedia category and Wikipedia article from all such pairs in Wikipedia, and manually investigated whether both category and article represented a person and whether they were linked by an *is-a* relation (positive:296, negative:704). Precision and recall of proposed method are 2.1 points and 11.8 points higher than those of Kobayashi's method, respectively. Among all 274,728 extracted as WPC-WPI pairs by our method, 269,233 was expected be correct.

	Precision	Recall	F-measure
Kobayashi’s method	95.9% (259/270)	87.5% (259/296)	91.5%
Proposed method	**98.0%** (294/300)	**99.3%** (294/296)	**98.7%**

Table 3: The extraction accuracy of the pairs of Wikipedia person category and person instance (WPC-WPI)

6 Discussions

We constructed a WPC hierarchy using the 8,357 categories created by combining extracted categories and training categories. The resulting WPC hierarchy has 224 root categories (Figure 4). Although the majority of the constructed ontology is interconnected, 194 person categories had no parent or child (2.3 % of all person categories). In rare cases, the category network has loops (e.g., “Historians” and “Scholars of history” are mutually interlinked).

Shibaki et al. (2009) presented a method for building a Japanese ontology from Wikipedia using *Goi-Taikei*, as its upper ontology. This method can create a single connected taxonomy with a single root category. We also hope to create a large-scale, single-root, and interconnected person ontology by using some upper ontology.

Our method is able to extract WPCs that do not match any *Goi-Taikei* instance (e.g., Violinists and Animators). Furthermore, our method is able to detect many ambiguous Wikipedia category labels correctly as person category. For example, “ファッションモデル (fashion model)” is ambiguous because the last word “モデル (model)” is ambiguous among three senses: person, artificial object, and abstract relation. Kobayashi’s method cannot extract a WPC if the last word of the category label does not match any instance in *Goi-Taikei*. Their method is error-prone if the last word has multiple senses in *Goi-Taikei* because it is based on simple pattern matching. Our method can handle unknown and ambiguous category labels since it uses machine learning-based classifiers whose features are extracted from neighboring categories.

Our method can extract *is-a* person category pairs that could not be extracted by Ponzetto et al. (2007) and Sakurai et al. (2008). Their methods use head matching in which a category link is labeled as an *is-a* relation only if the head words of category labels are matched. However, our method can extract *is-a* relations without reference to surface character strings, such as “ジャーナリスト(Journalists)” and “スポーツライター(Sports writers).” Among all 14,408 Wikipedia category pairs extracted as *is-a* relations in our method, 5,558 (38.6%) did not match their head words.

We investigated the learning curve of the machine learning-based classifier for extracting WPCs, in order to decide the appropriate amount of training data for future updates.

As we have already manually tagged all 39,767 Wikipedia categories, we randomly selected 30,000 categories and investigated the performance of our method when the number of the training data was changed from 1,000 to 30,000. The evaluation data was the remaining 9,767 categories.

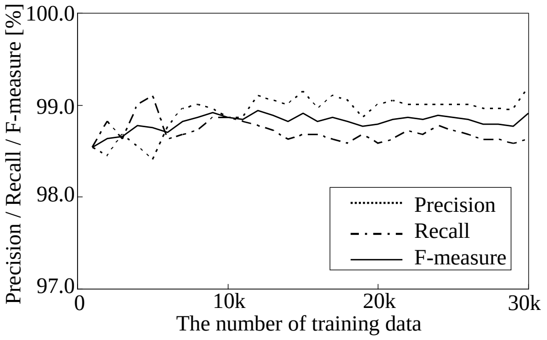


Figure 6: The effect of training data size to WPC extraction accuracy

Figure 6 shows the precision, recall, and F-measure for different training data sizes. F-measure differed only 0.4 points from 1,000 samples (98.5%) to 30,000 samples (98.9%). Figure 6 shows that the proposed method offers high accuracy in detecting WPCs with only a few thousand training examples.

Our method uses *similar categories* for creating features as well as the target Wikipedia category (Section 4.1). We compared the proposed method to a variant that does not use *similar categories* to confirm the effectiveness of this technique. Furthermore, our method uses the Japanese thesaurus, *Goi-Taikei*, to look up the semantic category of the words for creating the features for machine learning. We also compared the proposed method with the one that does not use semantic category (derived from *Goi-Taikei*) but instead uses word surface form for creating features (This one uses *similar categories*).

Figure 7 shows the performance of the classifiers for each type of features. We can clearly observe that using *similar categories* results in higher F-measure, regardless of the training data size. We also observe that when there is little training data, the method using word surface form as features results in drastically lower F-measures. In addition, its accuracy was consistently lower than the others even if the training data size was increased. Therefore, we can conclude that using *similar category* and *Goi-Taikei* are very important for creating good features for classification.

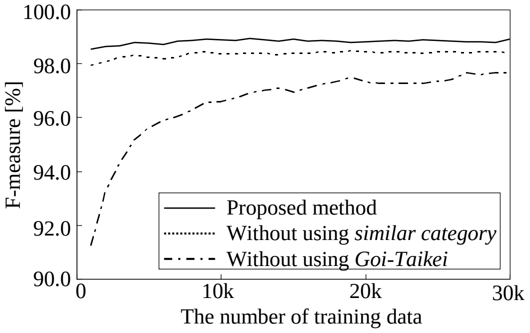


Figure 7: The effects of using *similar categories* and *Goi-Taikei*

In future, we will attempt to apply our method to other Wikipedia domains, such as organizations and products. We will also attempt to use other Japanese thesauri, such as Japanese WordNet. Furthermore, we hope to create a large-scale and single connected ontology. As a final note, we plan to open the person ontology constructed in this paper to the public on Web in the near future.

References

Bizer, C., J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. 2009. “DBpedia - A crystallization point for the web of data,” Web Semantics: Science, Services and Agents on the World Wide Web, vol. 7, No.3, pages 154-165.

Bond, Francis, Hitoshi Isahara, Kyoko Kanzaki, and Kiyotaka Uchimoto. 2008. Boot-strapping a wordnet using multiple existing wordnets. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 28-30.

Fellbaum, Christiane. 1998. WordNet: An Electronic Lexical Database, Language, Speech, and Communication Series. MIT Press.

Hearst, Marti A. 1992. Automatic acquisition of hyponyms from large text corpora. In Proceedings of the 14th Conference on Computational Linguistics (COLING), pages 539-545.

Ikehara, Satoru, Masahiro Miyazaki, Satoshi Shirai,Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Yoshifumi Ooyama, and Yoshihiko Hayashi, editors. 1997. *Nihongo Goi-Taikei – a Japanese Lexicon*. Iwanami Shoten. (in Japanese).

Kobayashi, Akio, Shigeru Masuyama, and Satoshi Sekine. 2008. A method for automatic construction of general ontology merging goitaikei and Japanese Wikipedia. In *Information Processing Society of Japan (IPSJ) SIG Technical Report2008-NL-187 (in Japanese)*, pages 7-14.

Ponzetto, S. P. and Michael Strube. 2007. Deriving a large scale taxonomy from Wikipedia. In *Proceedings of the 22nd Conference on the Advancement of Artificial Intelligence (AAAI)*, pages 1440–1445.

Sakurai, Shinya, Takuya Tejima, Masayuki Ishikawa, Takeshi Morita, Noriaki Izumi, and Takahira Yamaguchi. 2008. Applying Japanese Wikipedia for building up a general ontology. In *Japanese Society of Artificial Intelligence (JSAI) Technical Report SIG-SWO-A801-06 (in Japanese)*, pages 1-8.

Shibaki, Yumi, Masaaki Nagata and Kazuhide Yamamoto. 2009. Construction of General Ontology from Wikipedia using a Large-Scale Japanese Thesaurus. In *Information Processing Society of Japan (IPSJ) SIG Technical Report2009-NL-194-4. (in Japanese).*

Suchanek, Fabian M., Gjergji Kasneci, and GerhardWeikum. 2007. Yago: A core of semantic knowledge unifying wordnet and Wikipedia. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 697-706.

Sumida, Asuka, Naoki Yoshinaga, and Kentaro Torisawa. 2008. Boosting precision and recall of hyponymy relation acquisition from hierarchical layouts in Wikipedia. In *Proceedings of the Sixth Language Resources and Evaluation Conference(LREC)*, pages 28–30.

Using the Wikipedia Link Structure to Correct the Wikipedia Link Structure

Benjamin Mark Pateman
University of Kent
England
bmp7@kent.ac.uk

Colin Johnson
University of Kent
England
C.G.Johnson@kent.ac.uk

Abstract

One of the valuable features of any collaboratively constructed semantic resource (CSR) is its ability to – as a system – continuously correct itself. Wikipedia is an excellent example of such a process, with vandalism and misinformation being removed or reverted in astonishing time by a coalition of human editors and machine bots. However, some errors are harder to spot than others, a problem which can lead to persistent unchecked errors, particularly on more obscure, less viewed article pages. In this paper we discuss the problems of incorrect link targets in Wikipedia, and propose a method of automatically highlighting and correcting them using only the semantic information found in this encyclopaedia's link structure.

1 Introduction

Wikipedia, despite initial scepticism, is an incredibly robust semantic resource. Armed with a shared set of standards, legions of volunteers make positive changes to the pages of this vast encyclopaedia every day. Some of these editors may be casual – perhaps noticing an error in a page they were reading and being motivated to correct it – while others actively seek to improve the quality of a wide variety of pages that interest them. Facilitated by a relatively minimalist set of editing mechanics and incentives, Wikipedia has reached a state in which it is, for the most part, a reliable and stable encyclopaedia. Just enough regulation to prevent widespread vandalism or inaccuracy (including, on occasion, the temporary locking of particularly controversial pages), and enough editing freedom to maintain accuracy and relevance.

There are a number of potential approaches to minimizing misinformation and vandalism, falling into two broad categories: adding human incentives, and creating Wiki-crawling bots. There already exists a wide variety of natural and Wiki-based incentives (Kuznetsov, 2006) that have been crucial to the encyclopaedia's success. By implementing additional incentives, it may be possible to, for example, increase editor coverage of less-viewed articles. There are many avenues to explore regarding this, from additional community features such as a reputation system (Adler and de Alfaro, 2007), to ideas building upon recent work relating to games with a purpose (von Ahn, 2006), providing a form of entertainment that simultaneously aids page maintenance.

Wikipedia also benefits from a wide variety of bots and user-assistance tools. Some make the lives of dedicated editors easier (such as WikiCleaner[1]), providing an interface that facilitates the detection and correction of errors. Others carry out repetitive but important tasks, such as ClueBot[2], an anti-vandalism bot that reverts various acts of vandalism with surprising speed. Similar bots have been of great use in not only maintaining existing pages but also in adding new content (such as RamBot[3], a bot responsible for creating approximately 30,000 U.S city articles).

In recent years, researchers have taken an increasing interest in harnessing the semantic data contained in Wikipedia (Medelyan et al., 2009). To this end, the encyclopaedia now serves as not only a quick-lookup source for millions of people across the world, but also as an important semantic resource for a wide range of information retrieval, natural language processing and ontology building ap-

[1] https://launchpad.net/wikicleaner

[2] http://en.wikipedia.org/wiki/User:ClueBot

[3] http://en.wikipedia.org/wiki/User:Rambot

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 10–18,

plications. With all this utility, it is increasingly beneficial for Wikipedia to be as accurate and reliable as possible.

In this paper, we will discuss an algorithm that aims to use Wikipedia's inherent link structure to detect and correct errors within that very same structure. In Section 2 we will explore the nature and causes of this error, outlining the motivations for our algorithm. Section 3 discusses the inspirations for our approach, as well as our reasons for choosing it. We will then describe its method in detail, before evaluating its effectiveness and analysing its strengths and weaknesses.

2 A Reliable Encyclopaedia

"It's the blind leading the blind – infinite monkeys providing infinite information for infinite readers, perpetuating the cycle of misinformation and ignorance" (Keen, 2007). There has been much debate over the value of Wikipedia as a reliable encyclopaedia. Fallis (2008) talks at length about its epistemic consequences, acknowledging these criticisms but ultimately reaching a positive conclusion. In particular, he emphasizes the merits of Wikipedia in comparison with other easily accessible knowledge sources: If Wikipedia did not exist, people would turn to a selection of alternatives for quick-lookups, the collection of which are likely to be much less consistent, less verifiable and less correctable.

The fallacies of Wikipedia come from two sources: disinformation (an attempt to deceive or mislead) and misinformation (an honest mistake made by an editor). These can exist both in the textual content of an article, as well as the structural form of the encyclopaedia as a whole (e.g. the link structure or category hierarchy). The consequences can be measured in terms of the lifespan of such errors: a fairly harmless issue would be one that can be noticed and corrected easily, while those that are harder to detect and correct must be considered more troublesome.

For this reason, to be more potent on less frequently visited pages, as mentioned in Section 1. However, (Fallis, 2008) argues that "because they do not get a lot of readers, the potential epistemic cost of errors in these entries is correspondingly lower as well", suggesting that a balance is struck between misinformation and page traffic that stays somewhat consistent across all traffic levels. While inaccuracies may linger for longer on these less visited pages, it follows that fewer people are at risk of assuming false beliefs as a result.

An interesting pitfall of Wikipedia pointed out by Fallis (2008) comes as a result of the nature of its correctability. As readers of any piece of written information, certain factors can make us less trustworthy of its content; for example, grammatical or spelling mistakes, as well as blatant falsehoods. However, these are the first things to be corrected by Wikipedia editors, leaving what appears to be – on the surface – a credible article, but potentially one that embodies subtle misinformation that was not so quickly rectified.

2.1 Ambiguous Disambiguations

It is therefore important that methods of detecting and resolving the not-so-obvious inaccuracies are developed. One such not-so-obvious error can occur in Wikipedia's link structure. This problem stems from the polysemous nature of language (that is, that one word can map to multiple different meanings). In Wikipedia, different meanings of a word are typically identified by adding additional information in the relevant page's name. For example, the article "*Pluto (Disney)*" distinguishes itself from the article "*Pluto*" to avoid confusion between the *Disney* character and the dwarf planet. Adding extra information in brackets after the article name itself is Wikipedia's standard for explicitly disambiguating a word. Note that the article on the dwarf planet *Pluto* has no explicit disambiguation, because it is seen as the primary topic for this word. In other cases, no primary topic is assumed, and the default page for the word will instead lead directly to the disambiguation page (for example, see the Wikipedia page on "*Example*").

This system, while effective, is susceptible to human error when links are added or modified. The format for a link in WikiText is: "`[[PageName | AnchorText]]`" (the anchor text being optional). It is not hard to imagine, therefore, how a slightly careless editor might attempt to link to the article on *Pluto* (the *Disney* character) by typing "`[[Pluto]]`", assuming that this will link to the correct article, and not something completely different.

Is "*Jaguar*", generally the name of a fast feline, more likely to make you think of cars? "*Python*" is a genus of snake, but also a programming language to those involved in software development. Apple, a common fruit, but to a lot of people will be heavily associated with a well-known multinational corporation. These examples suggest that when a word takes on a new meaning, this new meaning – as long as it remains relevant – can become more recognizable than the original one (as yet another example, consider how your reaction to the word "*Avatar*" fluctuated in meaning as James Cameron's film went by). One particular potential problem is that someone editing an article will be focused on the context of that particular article, and will therefore be likely to not consider the polysemous nature of a word that they are using. For example, someone editing the article on the Apple *iPad* will have the company name Apple prominently in their mind, and therefore may momentarily forget about the existence of a particular kind of small round fruit.

The effects of these blunders can vary greatly depending on the word in question. For example, just about anyone who – expecting to be directed to a page on a *Disney* character – instead finds themselves at a page about a well-known dwarf planet in our Solar System, is going to know that there is an error in the source article. In this example, then, the error would be fixed very quickly indeed – faster still if the source page was popular (such as the article on *Disney* itself). However, there are cases where linking to the wrong sense of a polysemous word may not be as obvious an error for a lot of users. Someone following a link to "*Jagúar*" (the band) is less likely to notice a mistake if they're taken to the incorrect page of "*Jaguar (band)*" (a different band) than if they're taken to the incorrect page "*Jaguar*" (the feline). We argue that the extent of this problem depends on the difficulty of distinguishing between two different meanings of the same word. This difficulty is based upon two factors: the reader's level of background knowledge about the expected article, and the semantic similarity between it and the incorrect article being linked to. If the reader has absolutely no knowledge concerning the subject in question, they cannot be certain that they are viewing the correct page without further investigation. Furthermore, a reader with some relevant knowledge may still be unaware that they have been taken to the wrong page if the incorrectly linked-to page is semantically very similar to the page they were expecting. If these are common responses to a particular pair of polysemous articles, then it follows that a link error concerning them is likely to persist for longer without being corrected.

3 The Semantic Significance of Wikipedia's Link Structure

Wikipedia consists of, for the most part, unstructured text. Originally constructed with only the human user in mind, its design makes machine interpretations of its content difficult at best. However, the potential use of Wikipedia in a wide range of computational tasks has driven a strong research effort into ways of enriching and structuring its information to make it more suitable for these purposes. For example, DBpedia[4] takes data from Wikipedia and structures it into a consistent ontology, allowing all its information to be harnessed for various powerful applications, and is facilitating efforts towards realizing a semantic web (Bizer et al., 2009).

At the same time, research has also been carried out in ways of making use of the existing structure of Wikipedia for various natural language processing applications. For example, Shonhofen (2006) proposed using the hierarchical category structure of Wikipedia to categorize text documents. Another example of a system which makes use of word-sense disambiguation in the context of Wikipedia is the *Wikify!* system (Mihalcea and Csomai, 2007), which takes a piece of raw text and adds links to Wikipedia articles for significant terms. One of the biggest challenges for the authors of that system was linking to polysemous terms within the raw text. A combination of methods was used to determine the best disambiguation: overlap between concepts in the neighbourhood of the term and dictionary definitions of the various possible link targets, combined with a machine learning approach based on linguistic features.

In this paper we are concerned with another method of using Wikipedia without prior modifications: exploiting the nature of its network of links. This approach was pioneered by Milne and Witten (2007; 2008a; 2008b), responsible for developing the Wikipedia Link-Based Measure, an original measure of semantic relatedness that uses the unmodified network of links existing within Wikipedia.

Indeed, the link structure is one of the few elements of Wikipedia that can be easily interpreted by a machine without any restructuring. It contains

[4] http://dbpedia.org/

within it informal – often vague – relationships between concepts. Whereas, ideally, we would like to be dealing with labelled relationships, being able to directly analyse collections of untyped relationships is still very useful. Importantly, however, we must not concern ourselves with the significance of a single link (relationship), due to its class being unknown. In an article there may be links that are more significant – semantically speaking – than others, but this information cannot be retrieved directly. For example, the article on a famous singer might have a link to the village in which she grew up, but this is arguably – in most contexts – less semantically significant than the link to her first album, or the genre that describes her music.

Instead, then, we would like to look at collections of links, as these loosely summarize semantic information and de-emphasize the importance of knowing what type of relationship each link, individually, might express. Every single page on Wikipedia can be seen as a collection of links in this way; ignoring the raw, unstructured text within an article, we are still able to determine a great deal about its meaning just by looking at the underlying link structure. In doing this, comparing the similarity of two articles is as simple as comparing the outgoing links that each has. The more outgoing links that are common between the two articles, the more similar we can gauge them to be.

Looking at the links pointing to an article also provides us with additional cheap information. Of particular interest is deriving an estimated "commonness" of a concept by counting the number of links pointing in to it. The Wikipedia Link-Based Measure uses this information to weight each link, giving additional strength to links that have a lower probability of occurring. This accounts for the fact that two articles are less likely to share uncommon links; if they do, then this link overlap accounts for a higher degree of similarity. Conversely, two articles sharing a very common link (such as a page on a country or capital city) should not be considered very similar on that fact alone.

The motivations behind taking this approach for our link checking algorithm come largely from the inexpensive nature of this measure. While a large amount of potential information is ignored – such as the content of an article itself – the computational cost is an order of magnitude lower, and minimal preprocessing is required. With the English Wikipedia consisting of several million pages, and the search for incorrect links being essentially blind, processing speed is an important factor in providing useful page coverage.

4 Detecting Incorrect Links

The detection of incorrectly targeted links in Wikipedia is a trial of semantics; by estimating how similar in meaning a linked page is to the theme of an article, we can determine whether there might be an alternative page that would be more suitable. In finding significantly more suitable alternatives to these semantically unrelated links, we are able to hypothesise that the original link was incorrect. In the following subsections, we will describe the details of this algorithm.

4.1 Preparing the Database

Snapshots of Wikipedia can be downloaded from its database dump page[5], and then loaded into a local database. While this database is used by the algorithm, the practicality of such an application demands that live Wikipedia pages be used as the input. Checking a week old snapshot of Wikipedia for incorrect links will be less effective, as a number of them may well have been already fixed on the website itself. For this reason, the algorithm accepts a URL input of the page to be analysed, and will extract its current links directly.

4.2 Determining the Theme of an Article

The first step is to compute the semantic theme of the original article in question. This is done using an algorithm loosely based on that of Milne and Witten (2008a), which was discussed in section 3. To begin with, the original article is arranged as a list of linked pages (pages that it links directly to). Each of these derived pages is considered as a semantic "concept".

We represent each concept as a further list of its outgoing page links, creating a wide tree structure of depth 2, with the original article at the root (see Figure 1). The theme of this article is determined by propagating link information up the tree, essentially consolidating the individual themes of each of its concepts. As new links are discovered, they are assigned a commonness weighting (see section 3), and multiple encounters with the same link are tallied. For each link, this information (the common-

[5] http://dumps.wikimedia.org/backup-index.html

ness rating and link frequency) is used to sculpt the overall theme of the article.

4.3 Semantic Contribution

We use the phrase "semantic contribution" to describe how much meaning a particular concept "contributes" to the theme of the article in question. This is based on the nature of each of its links and how frequently they occur amongst the rest of the article's concepts. We therefore quantify the semantic contribution of a given concept by using the formula:

$$S_c = \sum_{l=1}^{n} \begin{cases} \log(f_l)w_l & \text{if } f_l >= 2 \\ 0 & \text{if } f_l < 2 \end{cases}$$

In other words, for each link l with a frequency (f, the number of times this link appears across all concepts) of 2 or more, its semantic contribution is a product of its frequency and its weight (w), as defined by:

$$w = \frac{1}{\log(i_l + 1)}$$

Where i_l is the total number of incoming links (Wikipedia-wide) pointing to the same target as link l. The total semantic contribution of a concept is the summation of all of the contributions of its outgoing links. By quantifying each concept in this manner, we can immediately see which concepts contribute a lot, and which contribute very little, to the theme of an article.

4.4 Extracting Dissimilar Links

With an aggregated theme established for an article, it is a simple task to flag up those concepts that have a low semantic contribution. Due to how semantic information was propagated up the tree (see the previous section), each concept represents some subset of the article's theme. Qualitatively speaking, this essentially equates to looking at how much of its theme overlaps with the most accentuated aspects of the article's theme. The dominant features of an article's theme will come from those links that are uncommon and frequently occurring, so any concept that consists of a good number of these links will be have a high semantic contribution.

By scoring each concept in terms of its contribution to the article theme, we are able to examine those concepts that scored particularly low. The value to use as a threshold for flagging potential errors is somewhat arbitrary, but in our experiments we have found best results using a simple variable threshold:

$$\text{Threshold} = \frac{\text{average contribution}}{2}$$

Any concepts with a semantic contribution below this value are considered as candidate errors, although it's important to note that, in many cases, a perfectly valid link can have a low contribution. For example, a link from a famous film director to a country he once filmed in. In these cases, however, we expect that it is unlikely for a more relevant alternative to be found.

4.5 Finding Possible Alternatives

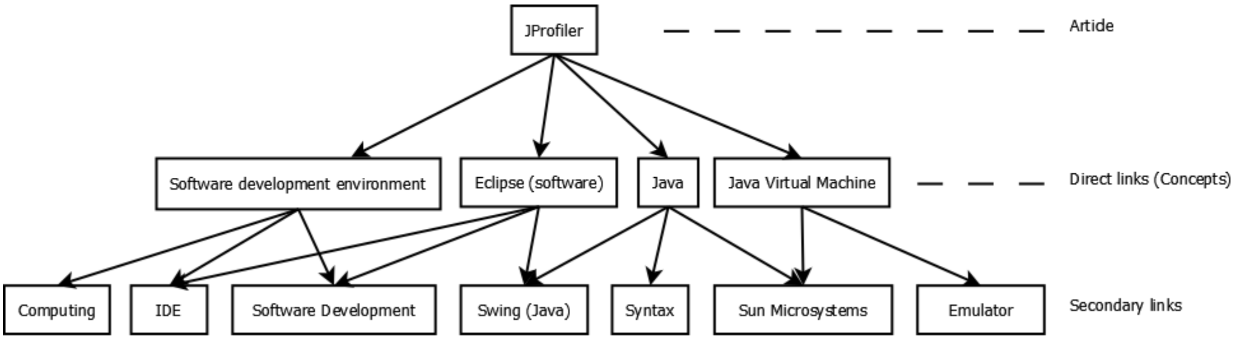
With one or more potentially incorrect links found, the algorithm must now search for alternative targets that are more suitable. This method is built on the assumption that the link is in error due to pointing towards the wrong disambiguation, accounting for the typical scenario of an editor linking to the wrong sense of a polysemous word.

An editor who has accidentally pointed to the article "*Pluto*" rather than "*Pluto (Disney)*" has not made any spelling errors. As we discussed in Section 2.1, the error is typically a result of a presumption being made on the most typical meaning of the target article. With this in mind, an error of this nature is likely to be resolved by looking at other articles that share the same name. There are a number of ways to do this, such as simply searching the database for all articles containing the word "*Pluto*". However, we chose instead to locate the relevant disambiguation page, if it exists (in this example, "*Pluto (disambiguation)*"). For the type of error we are targeting, this disambiguation page can be expected to contain the correct, intended page as one of its outgoing links.

4.6 Choosing the Best Alternative

With a list of possible alternatives for a particular weakly related concept, we then go about calculating their potential semantic contribution to the original article (using the same formula as was seen in section 4.4. To continue the example, the semantic contribution of "*Terry Pluto*" is unlikely to be at all high when considering the original article on *Disney*. The same goes for other possible alternatives, such as "*Pluto (newspaper)*" or "*Pluto Airlines*".

Figure 1: A simplified link structure diagram for the article on “*JProfiler*”.



However, the concept “*Pluto (Disney)*” contributes considerably more than the original link, and this becomes evidence to suggest it as a likely correction.

For each plausible alternative, a score is assigned based on the increased semantic contribution it provides over the original link. By doing this, the suggestions can be ordered from best to worst, expressing a degree of confidence in each option.

5 Evaluation

We evaluated the effectiveness of this algorithm by testing it on a snapshot of Wikipedia from November 2009. By using old Wikipedia pages we can, in most cases, easily validate our results against the now-corrected pages of live Wikipedia. However, finding examples of incorrectly linked articles is no simple task. Indeed, much of the justification for the algorithm this paper describes stems from the fact that finding these incorrect links is not easy, and actively searching for them is a somewhat tedious task. While we would like to leave our script crawling across Wikipedia detecting incorrect links by itself, in order to evaluate its performance we need to evaluate how well it performs on a set of pages that are known to contain broken links. It is impossible to generate such a set automatically, as by their nature these broken links are concerned with the meaning of the text on the pages.

We gauge the performance of our algorithm by looking at how many of the “best” suggestions (those with the highest calculated semantic contribution) given for a particular link are, in fact, correct.

5.1 Gathering Test Data

We found that a satisfactory method for finding incorrect links was to examine the incoming links pointing to a particularly ambiguous page. However, pages can have hundreds or thousands of incoming links, so we need to choose ones that are likely to be linked to in error, using ideas discussed in section 2.1. For example, if we look at the long list of links pointing towards the article “*Jaguar*”, we will mostly see articles relating to the animal: geographical locations, ecological information or pages concerning biology, for example. If, among these pages, we notice an out of place page – relating, perhaps, to cars, racing or business – we have reason to believe this article was supposed to be linking to something different (most likely, in this case, “*Jaguar Cars*”). After basic investigation we can confirm this and add it to our collection of pages for evaluation. While still not fast by any means, this method is considerably more effective than randomly meandering around the pages of Wikipedia in search of link errors. For this evaluation, we used the first 50 error-containing pages that were encountered using this method.

Another potentially effective method would be to download two chronologically separate snapshots of the Wikipedia database (for example, one taken a week before the other). We could then compare the incoming links to a particular article across both snapshots: If there are more incoming links in the newer snapshot than the old, then we can attempt to find them in the older snapshot and check their outgoing links. For example, the new snapshot might have a link from the article “*Jim Clark*” to “*Jaguar Cars*” that does not exist in the old snapshot. Upon checking the old snapshot’s version of the page on “*Jim Clark*”, we see it has a link to “*Jaguar*” and have immediately found a suitable error. This enables us to quickly find links that have been repaired in the time between the two snapshots, providing a fast, systematic method of gathering test data.

Nonetheless, finding a substantial set of examples of incorrectly linked pages is a significant challenge for work in this area. It is an important task, however, as without such a set it is impossible to determine a number of important features of a proposed correction algorithm. Firstly, without such a set it is impossible to determine which wrongly allocated links have been ignored by the algorithm, which is an important measure of the algorithm's success. Secondly, determining whether the algorithm has suggested the correct link requires that these correct links have been specified by a human user. As a result, the development of a substantial database of examples is an important priority for the development of this area of work.

5.2 Discussion

Overall, the results (given in Table 1) show that the algorithm performs well on this test set, with the best suggestion being the correct choice in 76.1% of cases.

As expected, the algorithm works best on larger articles with a well-established theme. For example, the articles on "*Austin Rover Group*" and "*CyberVision*" were riddled with links to incorrect pages, but with a total of 194 and 189 outgoing links respectively, there was sufficient information to confidently and accurately find the most suitable corrections, despite the number of errors. Conversely, "*Video motion analysis*", with only 7 outgoing links, fails to form a strong enough theme to even be able to highlight potential errors.

One might argue that the accurate result for the article on "*Synapse (disambiguation)*" is somewhat of an anomaly. Being a disambiguation page, there is inherently no common theme; typically, each link will point to a completely different area of semantic space. Correctly repairing the link to "*Java*" comes as somewhat of a coincidence, therefore, and it should be noted that disambiguation pages are not suited to this algorithm. Conversely, due to the nature of disambiguation pages, we might assume that users editing them are – in general – more careful about the target of their links, minimizing the occurrence of these sorts of errors.

There is a unique limitation with the algorithm that these results clearly highlight, however. An example of this lies in the results from programming-themed pages dealing with the link to "*Java*": There are a handful of recurring concepts being suggested, such as "*Java (programming language)*", "*Java (software platform)*" or "*Java Virtual Machine*". These suggestions are often accompanied by very similar values of semantic contribution, simply because they are all very semantically related to one another. As a result, if the theme of an article is related to one, it will be typically be related to them all. Which is the correct choice, if all are semantically valid? The one that fits best with the context of the sentence in which it is found.

This reveals an important limitation of this algorithm, in that the position of links within the text – and the surrounding text itself – is completely unknown to it. Dealing only with what is essentially a "bag of links", there is no information to discern which article (from a selection of strongly related articles) would be most appropriate for that particular link to point to. Indeed, in these isolated cases we observed the algorithms accuracy drop to 47%, although it should be noted that in almost all cases the correct link was suggested, just not as the best choice.

6 Conclusion

The results of our evaluation not only display the effectiveness of this algorithm at detecting and correcting typical link errors, but also clearly mark its limitations when dealing with multiple semantically similar suggestions. When considering the impact of these limitations, however, we must not forget that the algorithm was still able to recognize an invalid link, and was still able to offer the correct solution (often as the best choice). The impacts, then, are just on the consistency of the best choice being correct in these situations. However, the aim of this work was to build an algorithm that can be of significant assistance to a human editor's efficiency, and not to replace the editor. With that in mind, the output of the algorithm provides enough information to enable the editor to promptly pick the most appropriate suggestion, based on their own judgment.

While carrying out the evaluation on these 6 month old Wikipedia pages, we checked the results against the live pages of Wikipedia. A surprisingly large number (as many as 40%) of errors found had yet to be corrected half a year later, which, ulti-

Table 1: Counts of the correct link being given as the best suggestion.

Page Name	Best Correct	Page Name	Best Correct
Acropolis Rally	2/2	JProfiler	0/1
Austin Rover Group	6/6	KJots	0/1
Barabanki district	2/2	Lady G	0/1
Batch file	0/1	List of rapid application development tools	3/3
Belong (band)	1/1	Video motion analysis	0/1
Comparison of audio synthesis environments	3/4	Logo (programming language)	1/3
Comparison of network monitoring systems	2/3	Maria Jotuni	0/1
Computer-assisted translation	0/1	Mickey's delayed date	1/1
Convention over configuration	1/1	Neil Barret (Fashion Designer)	1/1
CyberVision	18/21	Ninja Gaiden (Nintendo Entertainment System)	2/3
Daimler 2.5 & 4.5 litre	1/2	Planetary mass	1/1
Dance music	3/3	Population-based incremental learning	1/1
Deiopea	1/1	Streaming Text Oriented Messaging Protocol	1/2
David Permut	3/3	Spiritual Warfare (video game)	1/2
Demon (video game)	1/1	Sonic Heroes	1/1
Disney dollar	1/1	Soulseek Records	2/2
DJ Hyper	1/1	Synapse (disambiguation)	1/1
DJ Qbert	1/2	Tellurium (software)	2/2
Eliseo Salazar	1/2	Testwell CTC++	1/1
Fixed point combinator	0/1	The Flesh Eaters (band)	3/3
Gravity Crash	1/1	Trans-Am Series	3/4
Hyphenation algorithm	2	Ultima IV: Quest of the Avatar	1/1
IBM Lotus Notes	1/2	Uma Thurman	4/6
Jaguar XFR	2/2	Unlabel	1/1
Jim Clark	0/1	Virtual World	1/2
		Total:	86/113

mately, is highly indicative of the potential benefits of this utility in repairing the errors that nobody knew existed.

7 Further Work

In continuing this work, there are a number of avenues to explore. Fundamentally, there is room to fine tune various aspects of the algorithm, such as the threshold value used to determine candidate errors, or the relationship between a link's frequency and its commonness. In doing so we might include additional variables, in particular investigating how the size of an article affects the algorithm, or the distribution of a central theme amongst its concepts.

Additionally, there is work to be done on constructing a practical application from this; adding, for example, an accessible GUI as well as direct Wikipedia integration to allow for users to easily commit corrected links to the Wikipedia server itself. This could lead to a further evaluation step in which we analyse the effectiveness of these corrections after the system has been running "in the wild" for a number of months. In order to use this system to correct the live Wikipedia it would be important to have an up-to-date local copy of Wikipedia in order to rapidly access the up-to-date link structure.

As mentioned earlier, an important challenge for the accurate evaluation of systems of this kind would be the development of a substantial, annotated database of examples of this kind of broken link. Clearly, it is difficult for a single development team to curate such a database, as the discovery process is time consuming. One approach to this would be through some form of crowdsourcing effort to gather a large number of examples. This could be as simple as encouraging readers of Wikipedia to report such corrections, for example by using a specific keyword in the revision notes made on that change. A more sophisticated approach could be to draw on the concept of *games with a purpose* (von Ahn, 2006), as exemplified by the *Google Image Labeler*[6] which uses a two-player game to find new tags for images. A game could be created based on the notion of presenting the user with a choice of links for a particular Wikipedia page, and rewarding them when they agree with another user on a target that is not currently pointed at by that link.

One further useful measure would be to devise a baseline algorithm to compare against. One possibility for this baseline would be to select the most heavily referenced choice from the list of candidates. This is similar to the approach used in data mining, where classifiers are compared against the naive classifier that classifies every instance as the most frequent item in the training set.

Finally, taking the reverse approach to the algorithm and looking primarily at incoming links – following the intuition behind our method of selecting test data (see section 5.1) – may prove very useful in locating articles that potentially contain incorrect links, allowing the algorithm to accurately and efficiently seek out pages to repair without having to crawl blindly across the entire encyclopaedia.

[6] http://images.google.com/imagelabeler/

References

Adler, B. Thomas and Luca de Alfaro. 2007. A content-driven reputation system for the wikipedia. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 261–270, New York, NY, USA. ACM.

Bizer, Christian, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. 2009. Dbpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, September.

Fallis, Don. 2008. Toward an epistemology of wikipedia. *Journal of the American Society for Information Science and Technology*, 59(10):1662–1674.

Keen, Andrew. 2007. *The Cult of the Amateur: How Today's Internet is Killing Our Culture*. Broadway Business, June.

Kuznetsov, Stacey. 2006. Motivations of contributors to wikipedia. *SIGCAS Comput. Soc.*, 36(2), June.

Medelyan, Olena, David Milne, Catherine Legg, and Ian H. Witten. 2009. Mining meaning from wikipedia. May.

Mihalcea, Rada and Andras Csomai. 2007. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 233–242, New York, NY, USA. ACM.

Milne, David and Ian H. Witten. 2008a. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artificial Intelligence*.

Milne, David and Ian H. Witten. 2008b. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA. ACM.

Milne, David. 2007. Computing semantic relatedness using wikipedia link structure. In *Proceedings of the New Zealand Computer Science Research Student Conference*.

Schonhofen, Peter. 2006. Identifying document topics using the wikipedia category network. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 456–462, Washington, DC, USA. IEEE Computer Society.

von Ahn, L. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Extending English ACE 2005 Corpus Annotation with Ground-truth Links to Wikipedia

Luisa Bentivogli
FBK-Irst
bentivo@fbk.eu

Pamela Forner
CELCT
forner@celct.it

Claudio Giuliano
FBK-Irst
giuliano@fbk.eu

Alessandro Marchetti
CELCT
amarchetti@celct.it

Emanuele Pianta
FBK-Irst
pianta@fbk.eu

Kateryna Tymoshenko
FBK-Irst
tymoshenko@fbk.eu

Abstract

This paper describes an on-going annotation effort which aims at adding a manual annotation layer connecting an existing annotated corpus such as the English ACE-2005 Corpus to Wikipedia. The annotation layer is intended for the evaluation of accuracy of linking to Wikipedia in the framework of a coreference resolution system.

1 Introduction

Collaboratively Constructed Resources (CCR) such as Wikipedia are starting to be used for a number of semantic processing tasks that up to few years ago could only rely on few manually constructed resources such as WordNet and SemCor (Fellbaum, 1998). The impact of the new resources can be multiplied by connecting them to other existing datasets, e.g. reference corpora. In this paper we will illustrate an on-going annotation effort which aims at adding a manual annotation layer connecting an existing annotated corpus such as the English ACE-2005 dataset[1] to a CCR such as Wikipedia. This effort will produce a new integrated resource which can be useful for the coreference resolution task.

Coreference resolution is the task of identifying which mentions, i.e. individual textual descriptions usually realized as noun phrases or pronouns, refer to the same entity. To solve this task, especially in the case of non-pronominal coreference, researchers have recently started to exploit semantic knowledge, e.g. trying to calculate the semantic similarity of mentions (Ponzetto and Strube, 2006) or their semantic classes (Ng, 2007; Soon et al., 2001). Up to now, WordNet has been one of the most frequently used sources of semantic knowledge for the coreference resolution task (Soon et al., 2001; Ng and Cardie, 2002). Researchers have shown, however, that WordNet has some limits. On one hand, although WordNet has a big coverage of the English language in terms of common nouns, it still has a limited coverage of proper nouns (e.g. Barack Obama is not available in the on-line version) and entity descriptions (e.g. president of India). On the other hand WordNet sense inventory is considered too fine-grained (Ponzetto and Strube, 2006; Mihalcea and Moldovan, 2001). In alternative, it has been recently shown that Wikipedia can be a promising source of semantic knowledge for coreference resolution between nominals (Ponzetto and Strube, 2006).

Consider some possible uses of Wikipedia. For example, knowing that the entity mention "Obama" is described on the Wikipedia page `Barack_Obama`[2], one can benefit from the Wikipedia category structure. Categories assigned to the `Barack_Obama` page can be used as semantic classes, e.g. "21st-century presidents of the United States". Another example of a useful Wikipedia feature are the links between Wikipedia pages. For instance, some Wikipedia pages contain links to the `Barack_Obama` page. Anchor texts of these links can provide alterna-

[1] `http://projects.ldc.upenn.edu/ace/`

[2] The links to Wikipedia pages are given displaying only the last part of the link which corresponds to the title of the page. The complete link can be obtained adding this part to `http://en.wikipedia.org/wiki/`.

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 19–27,

tive names of this entity, e.g. “Barack Hussein Obama” or “Barack Obama Junior”.

Naturally, in order to obtain semantic knowledge about an entity mention from Wikipedia one should link this mention to an appropriate Wikipedia page, i.e. to disambiguate it using Wikipedia as a sense inventory. The accuracy of linking entity mentions to Wikipedia is a very important issue. For example, such linking is a step of the approach to coreference resolution described in (Bryl et al., 2010). In order to evaluate this accuracy in the framework of a coreference resolution system, a corpus of documents, where entity mentions are annotated with ground-truth links to Wikipedia, is required.

The possible solution of this problem is to extend the annotation of entity mentions in a coreference resolution corpus. In the recent years, coreference resolution systems have been evaluated on various versions of the English Automatic Content Extraction (ACE) corpus (Ponzetto and Strube, 2006; Versley et al., 2008; Ng, 2007; Culotta et al., 2007; Bryl et al., 2010). The latest publicly available version is ACE 2005[3].

In this paper we present an extension of ACE 2005 non-pronominal entity mention annotations with ground-truth links to Wikipedia. This extension is intended for evaluation of accuracy of linking entity mentions to Wikipedia pages. The annotation is currently in progress. At the moment of writing this paper we have completed around 55% of the work. The extension can be exploited by coreference resolution systems, which already use ACE 2005 corpus for development and testing purposes, e.g. (Bryl et al., 2010). Moreover, English ACE 2005 corpus is multi-purpose and can be used in other information extraction (IE) tasks as well, e.g. relation extraction. Therefore, we believe that our extension might also be useful for other IE tasks, which exploit semantic knowledge.

In the following we start by providing a brief overview of the existing corpora annotated with links to Wikipedia. In Section 3 we describe some characteristics of the English ACE 2005 corpus, which are relevant to the creation of the extension. Next, we describe the general annotation principles and the procedure adopted to carry out the annotation. In Section 4 we present some analyses of the annotation and statistics about Inter-Annotator Agreement.

2 Related work

Recent approaches to linking terms to Wikipedia pages (Cucerzan, 2007; Csomai and Mihalcea, 2008; Milne and Witten, 2008; Kulkarni et al., 2009) have used two kinds of corpora for evaluation of accuracy: (i) sets of Wikipedia pages and (ii) manually annotated corpora. In Wikipedia pages links are added to terms “only where they are relevant to the context”[4]. Therefore, Wikipedia pages do not contain the full annotation of all entity mentions. This observation applies equally to the corpus used by (Milne and Witten, 2008), which includes 50 documents from the AQUAINT corpus annotated following the same strategy[5]. The corpus created by (Cucerzan, 2007) contains annotation of named entities only[6]. It contains 756 annotations, therefore for our purposes it is limited in terms of size.

Kulkarni et al. (2009) have annotated 109 documents collected from homepages of various sites with as many links as possible[7]. Their annotation is too extensive for our purposes, since they do not limit annotation to the entity mentions. To tackle this issue, one can use an automatic entity mention detector, however it is likely to introduce noise.

3 Creating the extension

The task consists of manually annotating the non-pronominal mentions contained in the English ACE 2005 corpus with links to appropriate Wikipedia articles. The objective of the work is to create an extension of ACE 2005, where all the mentions contained in the ACE 2005 corpus are disambiguated using Wikipedia as a sense repository to point to. The extension is intended for the

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T06

[4] http://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style

[5] http://www.nzdl.org/wikification/docs.html

[6] http://research.microsoft.com/en-us/um/people/silviu/WebAssistant/TestData/

[7] http://soumen.cse.iitb.ac.in/~soumen/doc/CSAW/

evaluation of accuracy of linking to Wikipedia in the framework of a coreference resolution system.

3.1 The English ACE 2005 Corpus

The English ACE 2005 corpus is composed of 599 articles assembled from a variety of sources selected from broadcast news programs, newspapers, newswire reports, internet sources and from transcribed audio. It contains the annotation of a series of entities (person, location, organization) for a total of 15,382 different entities and 43,624 mentions of these entities. A mention is an instance of a textual reference to an object, which can be either named (e.g. Barack Obama), nominal (e.g. the president), or pronominal (e.g. he, his, it). An entity is an aggregate of all the mentions which refer to one conceptual entity. Beyond the annotation of entities and mentions, ACE 05 contains also the annotation of local co-reference for the entities; this means that mentions which refer to the same entity in a document have been marked with the same ID.

3.2 Annotating ACE 05 with Wikipedia Pages

For the purpose of our task, not all the ACE 05 mentions are annotated, but only the named (henceforth NAM) and nominal (henceforth NOM) mentions. The resulting additional annotation layer will contain a total of 29,300 mentions linked to Wikipedia pages. As specifically regards the annotation of NAM mentions, information about local coreference contained in ACE 05 has been exploited in order to speed up the annotation process. In fact, only the first occurrence of the NAM mentions in each document has been annotated and the annotation is then propagated to all the other co-referring NAM mentions in the document.

Finally, it must be noted that in ACE 05, given a complex entity description, both the full extent of the mention (e.g. president of the United States) and its syntactic head (e.g. "president") are marked. In our Wikipedia extension only the head of the mention is annotated, while the full extent of the mention is available from the original ACE 05 corpus.

3.3 General Annotation Principles

Depending on the mention type to be annotated, i.e. NAM or NOM, a different annotation strategy has been followed. Each mention of type NAM is annotated with a link to a Wikipedia page describing the referred entity. For instance, "George Bush" is annotated with a link to the Wikipedia page `George_W._Bush`.

NOM mentions are annotated with a link to the Wikipedia page which provides a description of its appropriate sense. For instance, in the example "*I was driving Northwest of Baghdad and I bumped into these guys going around the capital*" the mention "capital" is linked to the page which provides a description of its meaning, i.e. `Capital_(political)`. Note that the object of linking is the textual description of an entity, and not the entity itself. In the example, even though from the context it is clear that the mention "capital" refers to Baghdad, we provide a link to the concept of capital and not to the entity Bagdad.

As a term can have both a more generic sense and a more specific one, depending on the context in which it occurs, mentions of type NOM can often be linked to more than one Wikipedia page. Whenever possible, the NOM mentions are annotated with a list of links to appropriate Wikipedia pages in the given context. In such cases, links are sorted in order of relevance, where the first link corresponds to the most specific sense for that term in its context, and therefore is regarded as the best choice. For instance, for the NOM mention head "President" which in the context identifies the United States President George Bush the annotation's purpose is to provide a description of the item "President", so the following links are selected as appropriate: `President_of_the_United_States` and `President`.

The correct interpretation of the term is strictly related to the context in which the term occurs. While performing the annotation, the context of the entire document has always been exploited in order to correctly identify the specific sense of the mention.

3.4 Annotation Procedure

The annotation procedure requires that the mention string is searched in Wikipedia in order to find the appropriate page(s) to be used for annotating the mention. In the annotation exercise, the annotators have always taken into consideration the context where a mention occurs, searching for both the generic and the most specific sense of the mention disambiguated in the context. In fact, in the example provided above, not only "President", but also "President of the United States" has been queried in Wikipedia as required by the context.

Not only the context, but also some features of Wikipedia must be mentioned as they affect the annotation procedure:

a. One element which contributes to the choice of the appropriate Wikipedia page(s) for one mention is the list of links proposed in Wikipedia's Disambiguation pages. Disambiguation pages are non-article pages which are intended to allow the user to choose from a list of Wikipedia articles defining different meanings of a term, when the term is ambiguous. Disambiguation pages cannot be used as links for the annotation as they are not suitable for the purposes of this task. In fact, the annotator's task is to disambiguate the meaning of the mention, so one link, pointing to a specific sense, is to be chosen. Disambiguation pages should always be checked as they provide useful suggestions in order to reach the appropriate link(s).

b. In the same way as Disambiguation pages, Wikitionary cannot be used as linking page, as it provides a list of possible senses for a term and not only one specific sense which is necessary to disambiguate the mention.

c. In Wikipedia, terms may be redirected to other terms which are related in terms of morphological derivation; i.e. searching for the term "Senator" you are automatically redirected to "Senate"; or querying "citizen" you are automatically redirected to "citizenship". Redirections have always been considered appropriate links for the term.

Some particular rules have been followed in order to deal with specific cases in the annotation, which are described below:

1. As explained before in Section 3.2, as a general rule the head of the ACE 05 mention is annotated with Wikipedia links. In those cases where the syntactic head of the mention is a multiword lexical unit, the ACE 05 practice is to mark as head only the rightmost item of the multiword. For instance, in the case of the multiword "flight attendant" only "attendant" is marked as head of the mention, although "flight attendant" is clearly a multiword lexical unit that should be annotated as one semantic whole. In our annotation we take into account the meaning of the whole lexical unit; so, in the above example, the generic sense of "attendant" has not been given, whereas `Flight_attendant` is considered as the appropriate link.

2. In some cases, in ACE 2005 pronouns like "somebody", "anybody", "anyone", "one", "others", were incorrectly marked as NOM (instead of PRO). Such cases, which amount to 117, have been marked with the tag "No Annotation".

3. When a page exists in Wikipedia for a given mention but not for the specific sense in that context the "Missing sense" annotation has been used. One example of "Missing sense" is for instance the term "heart" which has 29 links proposed in the "Disambiguation page" touching different categories (sport, science, anthropology, gaming, etc.), but there is no link pointing to the sense of "center or core of something"; so, when referring to the heart of a city, the term has been marked as "Missing sense".

4. When no article exists in Wikipedia for a given mention, the tag "No page" has been adopted.

5. Nicknames, i.e. descriptive names used in place of or in addition to the official name(s) of a person, have been treated as NAM. Thus, even if nicknames look like descriptions of individuals (and their reference should not be solved, following the general rule), they are actually used and annotated as

Number of annotated mentions	16310
Number of single link mentions	13774
Number of multi-link mentions	1458
Number of “No Page” annotations	481
Number of “Missing Sense” annotations	480
Number of “No Annotation” annotations	117
Total number of links	16851
Total number of links in multi-link mentions	3077

Table 1: Annotation data

proper names aliases. For example, given the mention “Butcher of Baghdad”, whose head “Butcher” is to be annotated, the appropriate Wikipedia link is `Saddam_Hussein`, automatically redirected from the searched string “Butcher of Baghdad”. The link `Butcher` is not appropriate as it provides a description of the mention. It is interesting the fact that Wikipedia itself redirects to the page of Saddam Hussein.

4 The ACE05-WIKI Extension

Up to now, the 55% of the markable mentions have been annotated by one annotator, amounting to 16,310 mentions. This annotation has been carried out by CELCT in a period of two months from February 22 to April 30, 2010, using the on-line version of Wikipedia, while the remaining 45% of the ACE mentions will be annotated during August 2010. The complete annotation will be freely available at: `http://www.celct.it/resources.php?id_page=acewiki2010`, while the ACE 2005 corpus is distributed by LDC[8].

4.1 Annotation Data Analysis

Table 1 gives some statistics about the overall annotation. In the following sections, mentions annotated with one link are called “single link”, whereas, mentions annotated with more than one link are named “multi-link”.

These data refer to the annotation of each single mention. It is not possible to give statistics at the entity level, as mentions have different ID depending on the documents they belong to, and the information about the cross-document co-reference is not available. Moreover, mentions of type NOM are annotated with different links depending on their disambiguated sense, making thus impossible to group them together.

Annotation	**Mention Type**	
	NAM	NOM
Single link mentions	6589	7185
Multi-link mentions	79	1379
Missing sense	96	384
No Page	440	41

Table 2: Distinction of NAM and NOM in the annotation

Most mentions have been annotated with only one link; if we consider multi-link mentions, we can say that each mention has been assigned an average of 2,11 links (3,077/1,458).

Data about “Missing sense” and “No page” are important as they provide useful information about the coverage of Wikipedia as sense inventory. Considering both “Missing sense” and “No page” annotations, the total number of mentions which have not been linked to a Wikipedia page amounts to 6%, equally distributed between “Missing sense” and “No page” annotations. This fact proves that, regarded as a sense inventory, Wikipedia has a broad coverage. As Table 2 shows, the mentions for which more than one link was deemed appropriate are mostly of type NOM, while NAM mentions have been almost exclusively annotated with one link only. The very few cases in which a NAM mention is linked to more than one Wikipedia page are primarily due to (i) mistakes in the ACE 05 annotation (for example, the mention “President” was erroneously marked as a NAM); (ii) or to cases where nouns marked as NAM could also be considered as NOMs (see for instance the mention “Marine”, to mean the Marine Corps).

Table 2 provides also statistics about the “Missing sense” and “No page” cases provided on mentions divided among the NAM and NOM type. The “missing sense” annotation concerns mostly the NOM category, whereas the NAM category is hardly affected. This attests the fact that persons, locations and organizations are well repre-

[8] `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T06`

sented in Wikipedia. This is mainly due to the encyclopedic nature of Wikipedia where an article may be about a person, a concept, a place, an event, a thing etc.; instead, information about nouns (NOM) is more likely to be found in a dictionary, where information about the meanings and usage of a term is provided.

4.2 Inter-Annotator Agreement

About 3,100 mentions, representing more than 10% of the mentions to be annotated, have been annotated by two annotators in order to calculate Inter-Annotator Agreement.

Once the annotations were completed, the two annotators carried out a reconciliation phase where they compared the two sets of links produced. Discrepancies in the annotation were checked with the aim of removing only the more rough errors and oversights. No changes have been made in the cases of substantial disagreement, which has been maintained.

In order to measure Inter-Annotator Agreement, two metrics were used: (i) the Dice coefficient to measure the agreements on the set of links used in the annotation[9] and (ii) two measures of agreement calculated at the mention level, i.e. on the group of links associated to each mention.

The Dice coefficient is computed as follows:

$$Dice = 2C/(A+B)$$

where C is the number of common links chosen by the two annotators, while A and B are respectively the total number of links selected by the first and the second annotator. Table 3 shows the results obtained both before and after the reconciliation process. Agreement before reconciliation is satisfactory and shows the feasibility of the annotation task and the reliability of the annotation scheme.

	BEFORE reconciliation	AFTER reconciliation
DICE	0.85	0.94

Table 3: Statistics about Dice coefficient

	BEFORE reconciliation	AFTER reconciliation
Complete	77.98%	91.82%
On first link	84.41%	95.58%

Table 4: Agreement at the mention level

Two measures of agreement at the mention level are also calculated. To this purpose, we count the number of mentions where annotators agree, as opposed to considering the agreement on each link separately. Mention-level agreement is calculated as follows:

$$\frac{\text{Number of mentions with annotation in agreement}}{\text{Total number of annotated mentions}}$$

We calculate both ”complete” agreement and agreement on the first link. As regards the first measure, a mention is considered in complete agreement if (i) it has been annotated with the same link(s) and (ii) in the case of multi-link mentions, links are given in the same order. As for the second measure, there is agreement on a mention if both the annotators chose the same first link (i.e. the one judged as the most appropriate), regardless of other possible links assigned to that mention. Table 4 provides data about both complete agreement and first link agreement, calculated before and after the annotators reconciliation.

4.3 Disagreement Analysis

Considering the 3,144 double-annotated mentions, the cases of disagreements amount to 692 (22,02%) before the reconciliation while they are reduced to 257 (8,18%) after that process. It is interesting to point out that the disagreements affect the mentions of type NOM in most of the cases, whereas mentions of type NAM are involved only in 3,8% of the cases.

Examining the two annotations after the reconciliation, it is possible to distinguish three kinds of disagreement which are shown in Table 5 to-

[9] The Dice coefficient is a typical measure used to compare sets in IR and is also used to calculate inter-annotator agreement in a number of tasks where an assessor is allowed to select a set of labels to apply to each observation. In fact, in these cases measures such as the widely used K are not good to calculate agreement. This is because K only offers a dichotomous distinction between agreement and disagreement, whereas what is needed is a coefficient that also allows for partial disagreement between judgments. In fact, in our case we often have a partial agreement on the set of links given for each mention. Also considering only the mentions for which a single link has been chosen, it is not possible to calculate K statistics in a straightforward way as the categories (i.e. the possible Wikipedia pages) in some cases cannot be determined a priori and are different for each mention. Due to these factors chance agreement cannot be calculated in an appropriate way.

Disagreement type	**Number of Disagreements**
1) No matching in the link(s) proposed	105 (40,85%)
2) No matching on the first link, but at least one of the other links is the same	14 (5,45%)
3) Matching on the first link and mismatch on the number of additional links	138 (53,70%)
Total Disagreements	257

Table 5: Types of disagreements

gether with the data about their distribution. An example of disagreement of type (1) is the annotation of the mention “crossing”, in the following context: *“Marines from the 1st division have secured a key Tigris River Crossing”*. Searching for the word “river crossing” in the Wikipedia searchbox, the Disambiguation Page is opened and a list of possible links referring to more specific senses of the term are offered, while the generic “river crossing” sense is missing. The annotators are required to choose just one of the possible senses provided and they chose two different links pointing to pages of more specific senses: {`Ford_%28river%29`} and {`Bridge`}.

Another example is represented by the annotation of the mention “area” in the context : *“Both aircraft fly at 125 miles per hour gingerly over enemy area”*. In Wikipedia no page exists for the specific sense of “area” appropriate in the context. Searching for “area” in Wikipedia, the page obtained is not suitable, and the Disambiguation page offers a list of various possible links to either more specific or more general senses of the term. One annotator judged the more general Wikipedia page `Area_(subnational_entity)` as appropriate to annotate the mention, while the second annotator deemed the page not suitable and thus used the “Missing sense” annotation.

Disagreement of type (2) refers to cases where at least one of the links proposed by the annotators is the same, but the first (i.e. the one judged as the most suitable) is different. Given the following context: *“Tom, You know what Liberals want”*, the two annotation sets provided for the mention “Liberal” are: {`Liberalism`} and {`Liberal_Party`, `Modern_liberalism_in_the_United_States`, `Liberalism`}.

The first annotator provided only one link for the mention “liberal”, which is different from the first link provided by second annotator. However, the second annotator provided also other links, among which there is the link provided by the first annotator.

Another example is represented by the annotation of the mention “killer”. Given the context: “*He’d be the 11th killer put to death in Texas*”, the two annotators provided the following link sets: {`Assassination`, `Murder`} and {`Murder`}. Starting from the Wikipedia disambiguation page, the two annotators agreed on the choice of one of the links but not on the first one.

Disagreement of type (3) refers to cases where both annotators agree on the first link, corresponding to the most specific sense, but one of them also added link(s) considered appropriate to annotate the mention. Given the context: “*7th Cavalry has just taken three Iraqi prisoners*”, the annotations provided for the term “prisoners” are: {`Prisoner_of_war`} and {`Prisoner_of_war`, `Incarceration`}. This happens when more than one Wikipedia pages are appropriate to describe the mention.

As regards the causes of disagreement, we see that the cases of disagreement mentioned above are due to two main reasons:

a. The lack of the appropriate sense in Wikipedia for the given mention

b. The different interpretation of the context in which the mention occurs.

In cases of type (a) the annotators adopted different strategies to perform their task, that is:

i. they selected a more general sense (i.e. “area” which has been annotated with `Area_(subnational_entity)`),

ii. they selected a more specific sense (see for example the annotations of the mentions “river crossing”).

iii. they selected the related senses proposed by the Wikipedia Disambiguation page (as in the annotation of “killer” in the example above).

Disagreement type (see above)	**Reas. a**	**Reas. b**	**Tot**
1) No match	95	10	105
2) No match on first link	4	10	14
3) Mismatch on additional links		138	138
Total	99 (38,5%)	158 (61,5%)	257

Table 6: Distribution of disagreements according to their cause

iv. they used the tag "Missing sense".

As Wikipedia is constantly evolving, adding new pages and consequently new senses, it is reasonable to think that the considered elements might find the appropriate specific/general link as time goes by.

Case (b) happens when the context is ambiguous and the information provided in the text allows different possible readings of the mention to be annotated, making thus difficult to disambiguate its sense. These cases are independent from Wikipedia sense repository but are related to the subjectivity of the annotators and to the inherent ambiguity of text.

Table 6 shows the distribution of disagreements according to their cause. Disagreements of type 1 and 2 can be due to both *a* and *b* reasons, while disagreements of type 3 are only due to *b*.

The overall number of disagreements shows that the cases where the two annotators did not agree are quite limited, amounting only to 8%. The analyses of the disagreements show some characteristics of Wikipedia considered as sense repository. As reported in Table 8, in the 61,5% of the cases of disagreement, the different annotations are caused by the diverse interpretation of the context and not by the lack of senses in Wikipedia. It is clear that Wikipedia has a good coverage and it proves to be a good sense disambiguation tool. In some cases it reveals to be too fine-grained and in other cases it remains at a more general level.

5 Conclusion

This paper has presented an annotation work which connects an existing annotated corpus such as the English ACE 2005 dataset to a Collaboratively Constructed Semantic Resource such as Wikipedia. Thanks to this connection Wikipedia becomes an essential semantic resource for the task of coreference resolution. On one hand, by taking advantage of the already existing annotations, with a relatively limited additional effort, we enriched an existing corpus and made it useful for a new NLP task which was not planned when the corpus was created. On the other hand, our work allowed us to explore and better understand certain characteristics of the Wikipedia resource. For example we were able to demonstrate in quantitative terms that Wikipedia has a very good coverage, at least as far as the kind of entity mentions which are contained in the ACE 2005 dataset (newswire) is concerned.

Acknowledgments

The research leading to these results has received funding from the ITCH project (`http://itch.fbk.eu`), sponsored by the Italian Ministry of University and Research and by the Autonomous Province of Trento and the Copilosk project (`http://copilosk.fbk.eu`), a Joint Research Project under Future Internet - Internet of Content program of the Information Technology Center, Fondazione Bruno Kessler.

We thank Giovanni Moretti from CELCT for technical assistance.

References

Bryl, Volha, Claudio Giuliano, Luciano Serafini, and Kateryna Tymoshenko. 2010. Using background knowledge to support coreference resolution. In *Proceedings of the 19th European Conference on Artificial Intelligence (ECAI 2010)*, August.

Csomai, Andras and Rada Mihalcea. 2008. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems*, 23(5):34–41.

Cucerzan, Silviu. 2007. Large-scale named entity disambiguation based on Wikipedia data. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, pages 708–716, Prague, Czech Republic, June. Association for Computational Linguistics.

Culotta, Aron, Michael L. Wick, and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 81–88.

Fellbaum, Christiane, editor. 1998. *WordNet: an electronic lexical database*. MIT Press.

Kulkarni, Sayali, Amit Singh, Ganesh Ramakrishnan, and Soumen Chakrabarti. 2009. Collective annotation of wikipedia entities in web text. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–466, New York, NY, USA. ACM.

Mihalcea, Rada and Dan I. Moldovan. 2001. Ez.wordnet: Principles for automatic generation of a coarse grained wordnet. In Russell, Ingrid and John F. Kolen, editors, *FLAIRS Conference*, pages 454–458. AAAI Press.

Milne, David and Ian H. Witten. 2008. Learning to link with wikipedia. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 509–518, New York, NY, USA. ACM.

Ng, Vincent and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111.

Ng, Vincent. 2007. Semantic class induction and coreference resolution. In *ACL 2007, Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics, June 23-30, 2007, Prague, Czech Republic*, pages 536–543.

Ponzetto, S. P. and M. Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, pages 192–199.

Soon, Wee Meng, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistic*, 27(4):521–544.

Versley, Yannick, Simone Paolo Ponzetto, Massimo Poesio, Vladimir Eidelman, Alan Jern, Jason Smith, Xiaofeng Yang, and Alessandro Moschitti. 2008. Bart: a modular toolkit for coreference resolution. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies*, pages 9–12.

Expanding textual entailment corpora from Wikipedia using co-training

Fabio Massimo Zanzotto
University of Rome ”Tor Vergata”
Rome, Italy
zanzotto@info.uniroma2.it

Marco Pennacchiotti
Yahoo! Lab
Sunnyvale, CA, 94089
pennac@yahoo-inc.com

Abstract

In this paper we propose a novel method to automatically extract large textual entailment datasets homogeneous to existing ones. The key idea is the combination of two intuitions: (1) the use of Wikipedia to extract a large set of textual entailment pairs; (2) the application of semi-supervised machine learning methods to make the extracted dataset homogeneous to the existing ones. We report empirical evidence that our method successfully expands existing textual entailment corpora.

1 Introduction

Despite the growing success of the Recognizing Textual Entailment (RTE) challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007), the accuracy of most textual entailment recognition systems are still below 60%. An intuitive way to improve performance is to provide systems with larger annotated datasets. This is especially true for machine learning systems, where the size of the training corpus is an important factor. As a consequence, several attempts have been made to train systems using larger datasets obtained by merging RTE corpora of different challenges. Unfortunately, experimental results show a significant decrease in accuracy (de Marneffe et al., 2006). There are two major reasons for this counter-intuitive result:

Homogeneity. As indicated by many studies (e.g. (Siefkes, 2008)), homogeneity of the training corpus is an important factor for the applicability of supervised machine learning models, since examples with similar properties often imply more effective models. Unfortunately, the corpora of the four RTE challenges are not homogenous. Indeed, they model different properties of the textual entailment phenomenon, as they have been created using slightly (but significantly) different methodologies. For example, part of the RTE-1 dataset (Dagan et al., 2006) was created using comparable documents, where positive entailments have a lexical overlap higher than negative ones (Nicholson et al., 2006; Dagan et al., 2006). Comparable documents have not been used as a source of later RTE corpora, making RTE-1 odd with respect to other datasets.

Corpus size. RTE corpora are relatively small in size (typically 800 pairs). The increase in size obtained by merging corpora from different challenges is not a viable solution. Much larger datasets, of one or more order of magnitude, are needed to capture the complex properties characterizing entailment.

A key issue for the future development of RTE is then the creation of datasets fulfilling two properties: (1) large size; (2) homogeneity wrt. existing RTE corpora. The task of creating large datasets is unfeasible for human annotators. Collaborative annotation environments such as the Amazon Mechanical Turk[1] can help to annotate pairs of sentences in positive or negative entailment (Zaenen, submitted; Snow et al., 2008). Yet, these environments can hardly solve the problem of finding relevant pairs of sentences. Completely automatic processes of dataset creation have been proposed (Burger and Ferro, 2005; Hickl et al., 2006). Unfortunately, these datasets are not homogeneous wrt. to the RTE datasets, as they are

[1] http://mturk.com

created using different methodologies. In this paper we propose a novel method to automatically extract entailment datasets which are guaranteed to be large and homogeneous to RTE ones. The key idea is the combination of two factors: (1) the use of Wikipedia as source of a large set of textual entailment pairs; (2) the application of semi-supervised machine learning methods, namely co-training, to make corpora homogeneous to RTE.

The paper is organized as follows. In Section 2 we report on previous attempts in automatically creating RTE corpora. In Section 3 we outline important properties that these corpora should have, and introduce our methodology to extract an RTE corpus from Wikipedia (the *WIKI corpus*), conforming to these properties. In Section 4 we describe how co-training techniques can be leveraged to make the WIKI corpus homogeneous to existing RTE corpora. In Section 5 we report empirical evidence that the combination of the WIKI corpus and co-training is successful. Finally, in Section 6 we draw final conclusions and outline future work.

2 Related Work

The first attempt to automatically create large RTE corpora was proposed by Burger and Ferro (Burger and Ferro, 2005), with the *MITRE corpus*, a corpus of positive entailment examples extracted from the XIE section of the Gigaword news collection (Graff, 2003). The idea of the approach is that the headline and the first paragraph of a news article should be (near-)paraphrase. Authors then collect paragraph-headline pairs as Text (T) - Hypothesis (H) examples, where the headlines plays the role of H. The final corpus consists of 100,000 pairs, with an estimated accuracy of 70% – i.e. two annotators checked a sample of about 500 pairs, and verified that 30% of these were either false entailments or noisy pairs. The major limitation of the Burger and Ferro (Burger and Ferro, 2005)'s approach is that the final corpus consist only of positive examples. Because of this imbalance, the corpus cannot be positively used by RTE learning systems.

Hickl et al. (2006) propose a solution to the problem, providing a methodology to extract both positive and negative pairs (the *LCC corpus*). A positive corpus consisting of 101,000 pairs is extracted similarly to (Burger and Ferro, 2005). Corpus accuracy is estimated on a sample of 2,500 examples, achieving 92% (i.e. almost all examples are positives), 22 points higher than Burger and Ferro. A negative corpus of 119,000 is extracted either: (1) selecting sequential sentences including mentions of a same named entity (98.000 pairs); (2) selecting pairs of sentences connected by words such as *even though, although, otherwise, but* (21,000 pairs). Estimated accuracy for the two techniques is respectively 97% and 94%.

Hickl and colleagues show that expanding the RTE-2 training set with the LCC corpus (the expansion factor is 125), their RTE system improves 10% accuracy. This suggests that by expanding with a large and balanced corpus, entailment recognition performance drastically improves. This intuition is later contradicted in a second experiment by Hickl and Bensley (2007). Authors use the LCC corpus with the RTE-3 training set to train a new RTE system, showing an improvement in accuracy of less than 1% wrt. the RTE-3 training alone.

Overall, evidence suggests that automatic expansion of the RTE corpora do not always lead to performance improvement. This highly depends on how balanced the corpus is, on the RTE system adopted, and on the specific RTE dataset that is expanded.

3 Extracting the WIKI corpus

In this section we outline some of the properties that a reliable corpus for RTE should have (Section 3.1), and show that a corpus extracted from Wikipedia conforms to these properties (Section 3.2).

3.1 Good practices in building RTE corpora

Previous work in Section 2 and the vast literature on RTE suggest that a "reliable" corpus for RTE should have, among others, the following properties:

(1) Not artificial. Textual entailment is a complex phenomenon which encompasses different linguistic levels. Entailment types range from very simple polarity mismatches and syntactic alternations, to very complex semantic and knowledge-

S'_1	*In this regard, some have charged the New World Translation Committee with being inconsistent.*
S'_2	*In this regard, some have charged the New World Translation Committee with not be consistent.*
S''_1	*The 'Stockholm Network' is Europe's only dedicated service organisation for market-oriented think tanks and thinkers.*
S''_2	*The 'Stockholm Network' is, according to its own site, Europe's only dedicated service organisation for market-oriented think tanks and thinkers.*

Figure 1: Sentence pairs from the Wikipedia revision corpus

based inferences. These different types of entailments are naturally distributed in texts, such as news and every day conversations. A reliable RTE corpus should preserve this important property, i.e. it should be rich in entailment types whose distribution in the corpus is similar to that in real texts; and should not include unrepresentative hand-crafted prototypical examples.

(2) Balanced and consistent. A reliable corpus should be *balanced*, i.e. composed by an equal or comparable number of positive and negative examples. This is particularly critical for RTE systems based on machine learning: highly imbalanced class distributions often result in poor learning performance (Japkowicz and Stephen, 2002; Kubat and Matwin, 1997). Also, the positive and negative subsets of the corpus should be *consistent*, i.e. created using the same methodology. If this property is not preserved, the risk is a learning system building a model which separates positive and negatives according to the properties characterizing the two methodologies, instead of those of the entailment phenomenon.

(3) Not biased on lexical overlap. A major criticism on the RTE-1 dataset was that it contained too many positive examples with high lexical overlap wrt. negative examples (Nicholson et al., 2006). Glickman et al. (2005) show that an RTE system using word overlap to decide entailment, surprisingly achieves an accuracy of 0.57 on RTE-1 test set. These performances are comparable to those obtained on the same dataset by more sophisticated and principled systems. Learning from this experience, a good corpus for RTE should avoid imbalances on lexical overlap.

(4) Homogeneous to existing RTE corpora. Corpus homogeneity is a key property for any machine learning approach (Siefkes, 2008). A new corpus for RTE should then model the same or similar entailments types of the reliable existing ones (e.g., those of the RTE challenges). If this is not the case, RTE system will be unable to learn a coherent model, thus resulting in a decrease in performance.

The MITRE corpus satisfies property (1), but does not (2) and (3), as it is highly imbalanced (it contains mostly positive examples), and is fairly biased on lexical overlap, as most examples of headline-paragraph pairs have many words in common. The LCC corpus suffers the problem of inconsistency, as positive and negative examples are derived with radically different methodologies. Both the MITRE and the LCC corpora are difficult to merge with the RTE challenge datasets, as they are not homogeneous – i.e. they have been built using very different methodologies.

3.2 Extracting the corpus from Wikipedia revisions

Our main intuition in using Wikipedia to build an entailment corpus is that the wiki framework should provide a natural source of non-artificial examples of true and false entailments, through its revision system. Wikipedia is an open encyclopedia, where every person can behave as an author, inserting new entries or modifying existing ones. We call *original entry* S_1 a piece of text in Wikipedia before it is modified by an author, and *revision* S_2 the modified text. The primary concern of Wikipedia authors is to reshape a document according to their intent, by adding or replacing pieces of text. Excluding vandalism, there are several reasons for making a revision: missing information, misspelling, syntactic errors, and, more importantly, disagreement on the content. For example, in Fig. 1, S''_1 is revised to S''_2, as the author disagrees on the content of S''_1.

Our hypothesis is that (S_1, S_2) pairs represent good candidates of both true and false entailment pairs (T, H), as they represent semantically close pieces of texts. Also, Wikipedia pairs conform to the properties listed in the previous section, as described in the following.

(S_1, S_2) pairs are *not artificial*, as we extract them from pieces of original texts, without any modification or post-processing. Also, pairs are rich of different entailment types, whose distribution is a reliable sample of language in use[2]. As shown later in the paper, a collection of (S_1, S_2) pairs is likely *balanced* on positive and negative examples, as authors either contradict the content of the original entry (false entailment) or add new information to the existing content (true entailment). Positive and negative pairs are guaranteed to be *consistent*, as they are drawn from the same Wikipedia source. Finally, the Wikipedia is *not biased in lexical overlap*: A sentence S_2 replacing S_1, usually changes only a few words. Yet, the meaning of S_2 may or may not change wrt. the meaning of S_1 – i.e. the lexical overlap of the two sentences is very high, but the entailment relation between S_1 and S_2 may be either positive or negative. For example, in Fig. 1 both pairs have high overlap, but the first is a positive entailment ($S_1' \rightarrow S_2'$), while the second is negative ($S_1'' \rightarrow S_2''$).

An additional interesting property of Wikipedia revisions is that the transition from S_1 to S_2 is commented by the author. The *comment* is a piece of text where authors explain and motivate the change (e.g. "general cleanup of spelling and grammar", "revision: Eysenck died in 1997!!"). Even if very small, the comment can be used to determine if S_1 and S_2 are in entailment or not. In the following section we show how we leverage comments to make the WIKI corpus *homogeneous* to those of the RTE challenges.

4 Expanding the RTE corpus with WIKI using co-training

Unlike the LCC corpus where negative and positive examples are clearly separated, the WIKI corpus mixes the two sets – i.e. it is unlabelled. In order to exploit the WIKI corpus in the RTE task, one should either manually annotate the corpus, or find an alternative strategy to leverage the corpus even if unlabelled. As manual annotation is unfeasible, we choose the second solution. The goal is then to expand a *labelled* RTE challenge training set with the *unlabelled* WIKI, so that the performance of an RTE system can increase over an RTE test set.

```
CO-TRAINING_ALGORITHM($L$,$U$,k)
returns $h_1$,$h_2$,$L_1$,$L_2$

    set $L_1 = L_2 = L$

    while stopping condition is not met

        – learn $h_1$ on $F_1$ from $L_1$, and learn $h_2$ on $F_1$ from $L_2$,
        – classify $U$ with $h_1$ obtaining $U_1$, and classify $U$ with $h_2$ obtaining $U_2$
        – select and remove $k$-best classified examples $u_1$ and $u_2$ from respectively $U_1$ and $U_2$
        – add $u_1$ to $L_2$ and $u_2$ to $L_1$
```

Figure 2: General co-training algorithm

In the literature, several techniques have been proposed to use unlabelled data to expand a training labelled corpus, e.g. Expectation-Maximization (Dempster et al., 1977). We here apply the co-training technique, first proposed by (Blum and Mitchell, 1998) and then successfully leveraged and analyzed in different settings (Abney, 2002). Co-training can be applied when the unlabelled dataset allows two independent views on its instances (*applicability condition*).

In this section, we first provide a short description of the co-training algorithm (Section 4.1). We then investigate if different RTE corpora conform to the applicability condition (Section 4.2). Finally, we show that our WIKI corpus conforms to the condition, and then apply co-training by creating two independent views (Section 4.3).

4.1 Co-training

The co-training algorithm uses unlabelled data to increase classification performance, and to indirectly increasing the size of labelled corpora. The algorithm can be applied only under a specific applicability condition: corpus' instances must have two *independent views*, i.e. they can be modeled by two independent feature sets.

We here adopt a slightly modified version of the

[2] It has been shown that web documents (as Wikipedia) are reliable samples of language (Keller and Lapata, 2003).

cotraining algorithm, as described in Fig.2. Under the applicability condition, instances are modeled on a feature space $F = F_1 \times F_2 \times C$, where F_1 and F_2 are the two independent views and C is the set of the target classes (in our case, true and false entailment). The algorithm starts with an initial set of training labelled examples L and a set of unlabelled examples U. The set L is copied in two sets L_1 and L_2, used to train two different classifiers h_1 and h_2, respectively using views F_1 and F_2. The two classifiers are used to classify the unlabelled set U, obtaining two different classifications, U_1 and U_2. Then comes the *co-training step*: the k-best classified instances in U_1 are added to L_2 and feed the learning of a new classifier h_2 on the feature space F_2. Similarly, the k-best instances in U_2 are added to L_1 and train a new classifier h_1 on F_1.

The procedure repeats until a stopping condition is met. This can be either a fixed number of added unlabelled examples (Blum and Mitchell, 1998), the performance drop on a control set of labelled instances, or a filter on the disagreement of h_1 and h_2 in classifying U (Collins and Singer, 1999). The final outcome of co-training is the new set of labelled examples $L_1 \cup L_2$ and the two classifier h_1 and h_2, obtained from the last iteration.

4.2 Applicability condition on RTE corpora

In order to leverage co-training for homogeneously expanding an RTE corpus, it is necessary to have a large unlabelled corpus which satisfies the applicability condition. Unfortunately, existing methodologies cannot guarantee the condition.

For example, the corpora from which the datasets of the RTE challenges were derived, were created from the output of applications performing specific tasks (e.g., Question&Answering, Information Extraction, Machine Translation, etc.). These corpora do not offer the possibility to create two completely independent views. Indeed, each extracted pair is composed only by the textual fragments of T and H, i.e. the only information available are the two pieces of texts, from which it is difficult to extract completely independent sets of features, as linguistic features tend to be dependent.

The MITRE corpus is extracted using two subsequent sentences, the title and the first paragraph. The LCC negative corpus is extracted using two correlated sentences or subsentences. Also in these two cases, it is very hard to find a view that is independent from the space of the sentence pairs.

None of the existing RTE corpora can then be used for co-training. In the next section we show that this is not the case for the WIKI corpus.

4.3 Creating independent views on the WIKI corpus

The WIKI corpus is naturally suited for co-training, as for each (S_1, S_2) pair, it is possible to clearly define two independent views:

- *content-pair view*: a set of features modeling the actual textual content of S_1 and S_2. This view is typically available also in any other RTE corpus.
- *comment view*: a set of features regarding the revision comment inserted by an author. This view represents "external" information (wrt. to the text fragments) which are peculiar of the WIKI corpus.

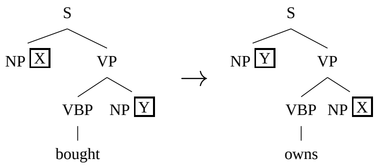
These two views are most likely independent. Indeed, the content-pair view deals with the content of the Wikipedia revision, while the comment view describes the reason why a revision has been made. This setting is very similar to the original one proposed for co-training by Blum and Mitchell (Blum and Mitchell, 1998), where the target problem was the classification of web pages, and the two independent views on a page were (1) its content and (2) its hyperlinks.

In the rest of this section we describe the feature spaces we adopt for the two independent views.

4.3.1 Content-pair view

The content-pair view is the classical view used in RTE. The original entry S_1 represents the Text T, while the revision S_2 is the Hypothesis H. Any feature space of those reported in the textual entailment literature could be applied. We here adopt the space that represents first-order syntactic rewrite rules (FOSR), as described in (Zanzotto and Moschitti, 2006). In this feature space, each feature represents a syntactic first-order or grounded rewrite rule. For example, the rule:

$\rho = l \rightarrow r =$



is represented by the feature $< l, r >$. A (T, H) pair activates a feature if it unifies with the related rule. A detailed discussion of the FOSR feature space is given in (Zanzotto et al., 2009) and efficient algorithms for the computation of the related kernel functions can be found in (Moschitti and Zanzotto, 2007; Zanzotto and Dell'Arciprete, 2009).

4.4 Comment view

A review comment is typically a textual fragment describing the reason why an author has decided to make a revision. In most cases the comment is not a well-formed sentence, as authors tend to use informal slang expressions and abbreviations (e.g. "details: Trelew Massacre; cat: Dirty War, copyedit", "removed a POV vandalism by Spylab", "dab ba:clean up using Project:AWB"). In these cases, where syntactic analysis would mostly fail, it is advisable to use simpler surface approaches to build the feature space. We then use a standard bag-of-word space, combined with a bag-of-2-grams space. For the first space we keep only meaningful content words, by using a standard stop-list including articles, prepositions, and very frequent words such as *be* and *have*. The second space should help in capturing small text fragments containing functional words: we then keep all words without using any stop-list.

5 Experiments

The goals of our experiments are the following: (1) check the quality of the WIKI corpus, i.e. if positive and negative examples well represent the entailment phenomenon; (2) check if WIKI contains examples similar to those of the RTE challenges, i.e. if the corpus is homogeneous to RTE; (3) check if the WIKI corpus improves classification performance when used to expand the RTE datasets using the co-training technique described in Section 4.

5.1 Experimental Setup

In order to check the above claims, we need to experiment with both manually labelled and unlabelled corpora. As unlabelled corpora we adopt:

wiki_unlabelled: An unlabelled WIKI corpus of about 3,000 examples. The corpus has been built by downloading 40,000 Wikipedia pages dealing with 800 entries about politics, scientific theories, and religion issues. We extracted original entries and revisions from the XML and wiki code, collecting an overall corpus of 20,000 (S_1, S_2) pairs. We then randomly selected the final 3,000 pairs.

news: A corpus of 1,600 examples obtained using the methods adopted for the LCC corpus, both for negative and positive examples (Hickl et al., 2006).[3] We randomly divided the corpus in two parts: 800 training and 800 testing examples. Each set contains an equal number of 400 positive and negative pairs.

As labelled corpora we use:

RTE-1,RTE-2, and **RTE-3**: The corpora from the first three RTE challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007). We use the standard split between training and testing.

wiki: A manually annotated corpus of 2,000 examples from the WIKI corpus. Pairs have been annotated considering the original entry as the H and the revision as T. Noisy pairs containing vandalism or grammatical errors were removed (these accounts for about 19% of the examples). In all, the annotation produced 945 positive examples (strict entailments and paraphrases) and 669 negative examples (reverse strict entailments and contradictions). The annotation was carried out by two experienced researchers, each one annotating half of the corpus. Annotation guidelines follow those used for the RTE challenges.[4]

[3] For negative examples, we adopt the headline - first paragraph extraction methodology.

[4] Annotators were initially trained on a small development corpus of 200 pairs. The inter-annotator agreement on this set, computed using the Kappa-statistics (Siegel and Castellan, 1988), was 0.60 corresponding to *substantial agreement*,

The corpus has been randomly split in three equally numerous parts: development, training, and testing. We kept aside the development to design the features, while we used training and testing for the experiments.

We use the Charniak Parser (Charniak, 2000) for parsing sentences, and SVM-light (Joachims, 1999) extended with the syntactic first-order rule kernels described in (Zanzotto and Moschitti, 2006; Moschitti and Zanzotto, 2007) for creating the FOSR feature space.

5.2 Experimental Results

The first experiment aims at checking the quality of the WIKI corpus, by comparing the performance obtained by a standard RTE system over the corpus in exam with those obtained over any RTE challenge corpus. The hypothesis is that if performance is comparable, then the corpus in exam has the same complexity (and quality) as the RTE challenge corpora. We then independently experiment with the $wiki$ and the $news$ corpora with the training-test splits reported in Section 5.1. As RTE system we adopt an SVM model learnt on the FOSR feature space described in Section 4.3.1.

The accuracies of the system on the $wiki$ and $news$ corpora are respectively 70.73% and 94.87%. The performance of the system on the $wiki$ corpus are in line with those obtained over the RTE-2 dataset (60.62%). This suggests that the WIKI corpus is at least as complex as the RTE corpora (i.e. positive and negatives are not trivially separable). On the contrary, the $news$ corpus is much easier to separate. Pilot experiments show that increasing the size of the $news$ corpus, accuracy reaches nearly 100%. This indicates that positive and negative examples in the $news$ corpus are extremely different. Indeed, as mentioned in Section 3.1, $news$ is not consistent – i.e. the extraction methods for the positives and the negatives are so different that the examples can be easily recognized using evidence not representative of the entailment phenomenon (e.g. for negative examples, the lexical overlap is extremely low wrt. positives).

Training Corpus	*Accuracy*
RTE-2	60.62
RTE-1	51.25
RTE-3	57.25
wiki	56.00
news	53.25
RTE-2+RTE-1	58.5
RTE-2+RTE-3	59.62
RTE-2+news	56.75
RTE-2+wiki	59.25
RTE-1+wiki	53.37
RTE-3+wiki	59.00

Table 1: Accuracy of different training corpora over RTE-2 test.

In a second experiment we aim at checking if WIKI is homogeneous to the RTE challenge corpora – i.e. if it contains (T, H) pairs similar to those of the RTE corpora. If this holds, we would expect the performance of the RTE system to improve (or at least not decrease) when expanding a given RTE challenge corpus with WIKI. de Marneffe et al. (2006) already showed in their experiment that it is extremely difficult to obtain better performance by simply expanding an RTE challenge training corpus with corpora of other challenges, since different corpora are usually not homogeneous.

We here repeat a similar experiment: we experiment with different combinations of training sets, over the same test set (namely, RTE-2 test). Results are reported in Table 1. The higher performance is the one of the system when trained on RTE-2 training set (second row) – i.e. a corpus completely homogeneous to RTE-2 would produce the same performance as RTE-2 training.

As expected, the models learnt on RTE-1 and RTE-3 perform worse (third and fourth rows): in particular, RTE-1 seems extremely different from RTE-2, as results show. The $wiki$ corpus is more similar to RTE-2 than the $news$ corpus, i.e. performance are higher. Yet, it is quite surprising that the $news$ corpus yields to a performance drop as in (Hickl et al., 2006) it shows a high performance increase.

The expansion of RTE-2 with the above corpora (seventh-tenth rows) lead to a drop in performance, suggesting that none of the corpora is completely homogeneous to RTE-2. Yet, the performance drop of the $wiki$ corpus (*RTE-2 +*

in line with the RTE challenge annotation efforts.

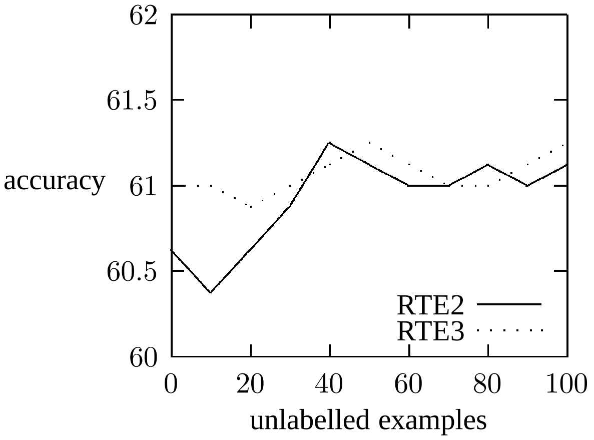


Figure 3: Co-training accuracy curve on the two corpora.

wiki) is comparable to the performance drop obtained using the other two RTE corpora (*RTE-2 + RTE-1* and *RTE-2 + RTE-3*). This indicates that $wiki$ is more homogeneous to RTE than $news$ – i.e. it contains (T, H) pairs that are similar to the RTE examples. Interestingly, $wiki$ combined with other RTE corpora (*RTE-1 + wiki* and *RTE-3 + wiki*) increases performance wrt. the models obtained with RTE-1 and RTE-3 alone (last two rows).

In a final experiment, we check if the WIKI corpus improves the performance when combined with the RTE-2 training in a co-training setting, as described in Section 4. This would confirm that WIKI is homogeneous to the RTE-2 corpus, and could then be successfully adopted in future RTE competitions. As test sets, we experiment both with RTE-2 and RTE-3 test. In the co-training, we use the RTE-2 training set as initial set L, and $wiki\_unlabelled$ as the unlabelled set U.[5]

Figure 3 reports the accuracy curves obtained by the classifier h_1 learnt on the content view, at each co-training iteration, both on the RTE-2 and RTE-3 test sets. As the comment view is not available in the RTE sets, the comment-view classifier become active only after the first 10 examples are fed as training from the content view classifier. As expected, performance increase for some steps and then become stable for RTE-3 and decrease for RTE-2. This is the only case in which we verified an increase in performance using corpora other than the official ones from RTE challenges. This result suggests that the WIKI corpus can successfully contribute to learn better textual entailment models for RTE.

6 Conclusions

In this paper we proposed a method for expanding existing textual entailment corpora that leverages Wikipedia. The method is extremely promising as it allows building corpora homogeneous to existing ones. The model we have presented is not strictly related to the RTE corpora. This method can then be used to expand corpora such as the Fracas test-suite (Cooper et al., 1996) which is more oriented to specific semantic phenomena.

Even if the performance increase of the completely unsupervised cotraining method is not extremely high, this model can be used to semi-automatically expanding corpora by using active learning techniques (Cohn et al., 1996). The initial increase of performances is an interesting starting point.

In the future, we aim at releasing the annotated portion of the WIKI corpus to the community; we will also carry out further experiments and refine the feature spaces. Finally, as Wikipedia is a multilingual resource, we will use the WIKI methodology to semi-automatically build RTE corpora for other languages.

References

Steven Abney. 2002. Bootstrapping. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 360–367, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, and Idan Magnini, Bernardo Szpektor. 2006. The second pascal recognising textual entailment challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *COLT: Proceedings of the Conference on Computational Learning Theory*. Morgan Kaufmann.

[5] Note that only $wiki\_unlabelled$ allows both views described in Section 4.3.

John Burger and Lisa Ferro. 2005. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 49–54, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Eugene Charniak. 2000. A maximum-entropy-inspired parser. In *Proc. of the 1st NAACL*, pages 132–139, Seattle, Washington.

David A. Cohn, Zoubin Ghahramani, and Michael I. Jordan. 1996. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145.

Michael Collins and Yoram Singer. 1999. Unsupervised models for named entity classification. In *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, pages 100–110.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Johan Van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, and Steve Pulman. 1996. Using the framework. Technical Report LRE 62-051 D-16, The FraCaS Consortium. Technical report.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In Quionero-Candela et al., editor, *LNAI 3944: MLCW 2005*, pages 177–190, Milan, Italy. Springer-Verlag.

Marie-Catherine de Marneffe, Bill MacCartney, Trond Grenager, Daniel Cer, Anna Rafferty, and Christopher D. Manning. 2006. Learning to distinguish valid textual entailments. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

A.P. Dempster, N.M. Laird, and D.B. Rubin. 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague, June. Association for Computational Linguistics.

Oren Glickman, Ido Dagan, and Moshe Koppel. 2005. Web based probabilistic textual entailment. In *Proceedings of the 1st Pascal Challenge Workshop*, Southampton, UK.

David Graff. 2003. English gigaword.

Andrew Hickl and Jeremy Bensley. 2007. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 171–176, Prague, June. ACL.

Andrew Hickl, John Williams, Jeremy Bensley, Kirk Roberts, Bryan Rink, and Ying Shi. 2006. Recognizing textual entailment with LCCs GROUNDHOG system. In *Proceedings of the 2nd PASCAL Challenge Workshop on RTE*, Venice, Italy.

N. Japkowicz and S. Stephen. 2002. The class imbalance problem: A systematic study. *Intelligent Data Analysis*, 6(5).

Thorsten Joachims. 1999. Making large-scale svm learning practical. In B. Schlkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods-Support Vector Learning*. MIT Press.

Frank Keller and Mirella Lapata. 2003. Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3), September.

M. Kubat and S. Matwin. 1997. Addressing the curse of inbalanced data sets: One-side sampleing. In *Proceedings of the 14th International Conference on Machine Learning*, pages 179–186. Morgan Kaufmann.

Alessandro Moschitti and Fabio Massimo Zanzotto. 2007. Fast and effective kernels for relational learning from texts. In *Proceedings of the International Conference of Machine Learning (ICML)*, Corvallis, Oregon.

Jeremy Nicholson, Nicola Stokes, and Timothy Baldwin. 2006. Detecting entailment using an extended implementation of the basic elements overlap metric. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Christian Siefkes. 2008. *An Incrementally Trainable Statistical Approach to Information Extraction*. VDM Verlag, Saarbrucken, Germany.

S. Siegel and Jr. N. J. Castellan. 1988. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on EmNLP*, pages 254–263, Honolulu, Hawaii. ACL.

Annie Zaenen. submitted. Do give a penny for their thoughts. *Journal of Natural Language Engineering*.

Fabio Massimo Zanzotto and Lorenzo Dell'Arciprete. 2009. Efficient kernels for sentence pair classification. In *Conference on Empirical Methods on Natural Language Processing*, pages 91–100, 6-7 August.

Fabio Massimo Zanzotto and Alessandro Moschitti. 2006. Automatic learning of textual entailments with cross-pair similarities. In *Proceedings of the 21st Coling and 44th ACL*, pages 401–408, Sydney, Australia, July.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. 2009. A machine learning approach to textual entailment recognition. *NATURAL LANGUAGE ENGINEERING*, 15-04:551–582. Accepted for pubblication.

Pruning Non-Informative Text Through Non-Expert Annotations to Improve Aspect-Level Sentiment Classification

Ji Fang
Palo Alto Research Center
`Ji.Fang@parc.com`

Bob Price
Palo Alto Research Center
`Bob.Price@parc.com`

Lotti Price
Palo Alto Research Center
`Lotti.Price@parc.com`

Abstract

Sentiment analysis attempts to extract the author's sentiments or opinions from unstructured text. Unlike approaches based on rules, a machine learning approach holds the promise of learning robust, high-coverage sentiment classifiers from labeled examples. However, people tend to use different ways to express the same sentiment due to the richness of natural language. Therefore, each sentiment expression normally does not have many examples in the training corpus. Furthermore, sentences extracted from unstructured text (e.g., I filmed my daughter's ballet recital and could not believe how the auto focus kept blurring then focusing) often contain both informative (e.g., the auto focus kept blurring then focusing) and extraneous non-informative text regarding the author's sentiment towards a certain topic. When there are few examples of any given sentiment expression, extraneous non-sentiment information cannot be identified as noise by the learning algorithm and can easily become correlated with the sentiment label, thereby confusing sentiment classifiers. In this paper, we present a highly effective procedure for using crowd-sourcing techniques to label informative and non-informative information regarding the sentiment expressed in a sentence. We also show that pruning non-informative information using non-expert annotations during the training phase can result in classifiers with better performance even when the test data includes non-informative information.

1 Introduction

Noise in training data can be derived either from noisy labeling or from noisy features. It has been shown that labeling quality is one of the important factors that impacts the performance of a learned model, and that this quality can be improved by approaches such as using multiple labelers (Sheng et al., 2008). However, noisy features can be an inherent characteristic for some text mining tasks, and it is unclear how they should be handled.

For example, sentiment analysis/opinion mining from unstructured user generated content such as online reviews and blogs often relies on learning sentiments from word-based features extracted from the training sentences and documents (Pang et al., 2002; Dave et al., 2003; Kim and Hovy, 2005). However, not all words in the training data carry information about sentiment. For example, in sentence (1),

(1)*I filmed my daughter's ballet recital and could not believe how the auto focus kept blurring then focusing.*

although words such as *auto focus, blurring* and *focusing* are informative for learning sentiment regarding the auto focus capability of the camera, words such as *film, daughter* and *ballet recital* are not informative for that type of sentiment, and they form noise if included as training data.

If the training data contain a lot of examples such as (2) in which words such as *film, daughter* and *ballet recital* also appear, but the sentence is not labelled as invoking sentiment regarding auto focus, a machine learning algorithm might learn

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 37–45,

that such words are not informative for sentiment classification.

(2)*I filmed my daughter's ballet recital and could not believe how good the picture quality was.*

However, due to the richness of natural language, people tend to use different ways to describe a similar event or to express a similar opinion. Consequently, repeated use of the same expression is not common in the training data for sentiment classification. Note that this difficulty cannot be simply overcome by increasing the size of the training data. For example, a search on the completely natural phrase "I filmed my daughter's ballet recital" in Google and Bing returns the same exact sentence as shown in (1). In other words, there appears to be only one sentence containing that exact phrase, which implies that even if we use the entire web as our training data set we would not find an example such as (2) to help the learning algorithm to determine which feature words in (1) are informative and which are not. Therefore, data sparsity is an inherent problem for a task such as sentiment analysis, and if we adopt the bag-of-words approach for sentiment classification (Pang et al., 2002), which uses the words that appear in sentences as training features, our training data will unavoidably include many noisy non-informative features.

This paper presents a crowd-sourcing technique to identify and prune the non-informative features. We explore the effect of using non-expert annotations to gain low-noise training data for sentiment classification. We show that the cleaner training data obtained from non-expert annotations significantly improve the performance of the sentiment classifier. We also present evidence that this improvement is due to reduction in confusion between classes due to noise words.

The remainder of this paper is organized as follows. Section 2 discusses the related work. Section 3 describes our approach for pruning non-informative features. Section 4 presents an empirical study on the effect of training on informative features in the domain of sentiment analysis. Conclusions are summarized in Section 5.

2 Related Work

Feature selection in the domain of sentiment analysis has focused on the following issues.

a) Should word-based features be selected based on frequency or presence?

It has been shown that compared to word frequency, word presence is a better sentiment indicator (Pang et al., 2002; Wiebe et al., 2004; Yang et al., 2006). In other words, unlike in other domains such as topic classification where the frequency of words provides useful information regarding the topic class, sentiment information is not normally indicated by the frequency of certain words, because people are unlikely to repeatedly use the same word or phrase to express an opinion in one document. Instead, Researchers (Pang et al., 2002) found that selecting features based on word presence rather than word frequency leads to better performance in the domain of sentiment analysis.

b) Which are more useful features: unigrams, higher-order n-grams or syntactically related terms?

This issue seems to be debatable. While some researchers (Pang et al., 2002) reported that unigrams outperform both bigrams as well as the combination of unigrams and bigrams in classifying movie reviews based on sentiment polarity, some others (Dave et al., 2003) reported the opposite in some settings.

Similarly, some (Dave et al., 2003) found syntactically related terms are not helpful for sentiment classification, whereas others (Gamon, 2004; Matsumoto et al., 2005; Ng et al., 2006) found the opposite to be true.

c) In terms of part-of-speech, which types of words are more useful features?

Adjectives and adverbs are commonly used as features for sentiment learning (Mullen and Collier, 2004; Turney, 2002; Whitelaw et al., 2005). However, more recent studies show that all content words including nouns, verbs, adjectives and adverbs are useful features for sentiment analysis (Dillard, 2007).

Regardless of which types of features are used, these traditional approaches are still inherently noisy in the sense that non-informative words/features within each sentence are included as described in Section 1. As far as we are aware, this is an issue that has not been addressed.

The closest works are Riloff et al. (Riloff and Wiebe, 2003) and Pang et al. (Pang et al., 2002)'s work. Riloff et al. explored removing the features that are subsumed in other features when a combination of different types of features such as unigrams, bigrams and syntactically related terms is used. Pang et al. speculated that words that appear at certain positions in a movie review are more informative for the overall opinion reflected in that review. However, according to Pang et al., for the task of predicting the overall polarity of a movie review, training on word features assumed to be more informative resulted in worse performance than training on all word features appearing in the reviews.

Our approach is different in that we try to identify and prune non-informative word features at the sentence level. We focus on identifying which portion of the sentence is informative for sentiment classification. We then completely remove the non-informative portion of the sentence and prevent any terms occurring in that portion from being selected as feature vectors representing that sentence. Note that the classification of words as non-informative is not related to their positions in a sentence nor to their frequency count in the training corpus. Instead, whether a word is informative depends purely on the semantics and the context of the sentence. For example, the word *big* would be non-informative in (3), but informative in (4).

(3)*That was a big trip, and I took a lot of pictures using this camera.*

(4)*This camera has a big LCD screen.*

Unlike the traditional approach of using expert annotation to identify the non-informative text in a sentence, we instead use non-expert annotations without external gold standard comparisons. There have been an increasing number of experiments using non-expert annotations for various Natural Language Processing (NLP) tasks. For example, Su et al. (Su et al., 2007) use non-expert annotations for hotel name entity resolution. In (Nakov, 2008), non-expert annotators generated paraphrases for 250 noun-noun compounds, which were then used as the gold standard data for evaluating an automatic paraphrasing system. Kaisser and Lowe (Kaisser and Lowe, 2008) also use non-experts to annotate answers contained in sentences and use the annotation results to help build a question answering corpus. Snow et al. (Snow et al., 2008) reported experiments using non-expert annotation for the following five NLP tasks: affect recognition, word similarity, recognizing textual entailment, event temporal ordering, and word sense disambiguation.

This paper presents a study of using non-expert annotations to prune non-informative word features and training a sentiment classifier based on such non-expert annotations. The following section describes our approach in detail.

3 Non-Informative Feature Pruning Through Non-Expert Annotations

To prune the non-informative features, a traditional approach would be to hire and train annotators to label which portion of each training sentence is informative or non-informative. However, this approach is both expensive and time consuming. We overcome these issues by using crowdsourcing techniques to obtain annotations from untrained non-expert workers such as the ones on the Amazon Mechanical Turk (AMT) platform[1]. To illustrate our approach, we use an example for sentiment analysis below.

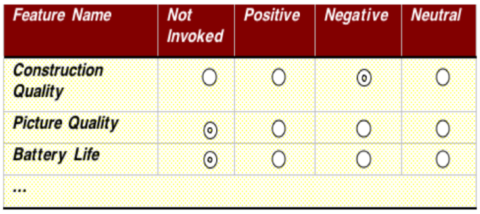
The key to our approach relies on careful design of simple tasks or HITs that can elicit the necessary information for both labeling the sentiment information and pruning the non-informative text of a sentence. These tasks can be performed quickly and inexpensively by untrained non-expert workers on the AMT platform. We achieved this goal by designing the following two experiments.

Experiment 1 asks the workers to judge whether a sentence indicates an opinion towards a certain aspect of the camera, and if so, whether the opinion is positive, negative or neutral. For example, the proper annotations for sentence (5) would be as shown in Figure 1.

[1] This is an online market place that offers a small amount of money to people who perform some "Human Intelligence Tasks" (HITs). `https://www.mturk.com/mturk/welcome`

(5) *On my trip to California, the camera fell and broke into two pieces.*

Figure 1: Experiment 1



Feature Name	Not Invoked	Positive	Negative	Neutral
Construction Quality	○	○	◉	○
Picture Quality	◉	○	○	○
Battery Life	◉	○	○	○
...				

We randomly selected 6100 sentences in total for this experiment from the Multi-Domain Sentiment Dataset created by Blitzer et al. (Blitzer et al., 2007). Each sentence was independently annotated by two AMT workers. Each annotation consisted of a sentence labeled with a camera aspect and a sentiment toward that aspect.

One unique characteristic of Experiment1 is that it makes the detection of unreliable responses very easy. Because one sentence is unlikely to invoke many different aspects of cameras, an annotation is thus suspicious if many aspects of camera are annotated as being invoked. Figure 2 and Figure 3 illustrate the contrast between a normal reliable response and a suspicious unreliable response.

Due to this favorable characteristic of Experiment 1, we did not have to design a qualification test. We approved all of the assignments; however we later filtered out the detected suspicious responses, which accounted for 8% of the work. Even though we restricted our AMT workers to those who have an approval rate of 95% or above, we still found 20% of them unreliable in the sense that they provided suspicious responses.

Given our ability to detecting suspicious responses, we believe it is very unlikely for two reliable AMT workers to annotate any given sentence exactly the same way merely by chance. Therefore, we consider an annotation to be gold when both annotators marked the same sentiment toward the same aspect. We obtained 2718 gold-standard annotations from the reliable responses. We define the agreement rate of annotations as follows.

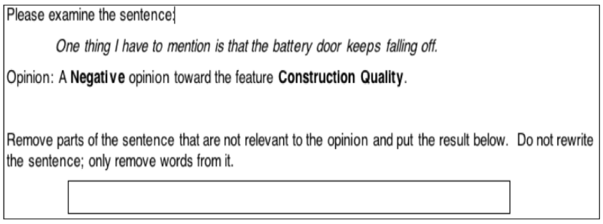
$$AgreementRate = \frac{NumberofGoldAnnotations \times 2}{TotalNumberofAnnotations}. \quad (1)$$

Based on this measure, the agreement rate of the AMT workers in this study is 48.4%.

We held randomly selected 587 gold annotated sentences as our test set, and used the remaining 2131 sentences as our training sentences. To prune the non-informative text from the training sentences, we put the 2131 sentences through Experiment 2 as described below.

Experiment 2 asks the workers to point out the exact portion of the sentence that indicates an opinion. The opinion and its associated feature name are displayed along with the sentence in which they appear. Such information is automatically generated from the results derived from Experiment 1. An example of Experiment 2 is given in Figure 4.

Figure 4: Experiment 2



Please examine the sentence:

One thing I have to mention is that the battery door keeps falling off.

Opinion: A **Negative** opinion toward the feature **Construction Quality**.

Remove parts of the sentence that are not relevant to the opinion and put the result below. Do not rewrite the sentence; only remove words from it.

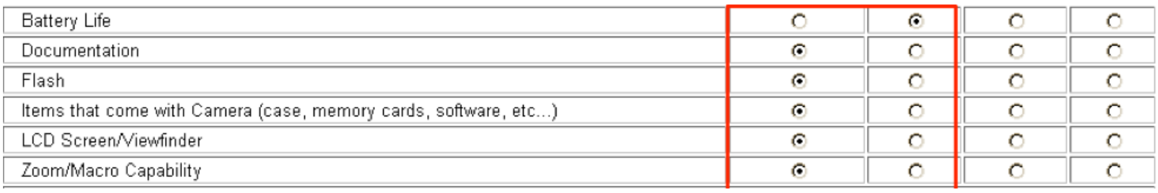
The expected answer for this example is *the battery door keeps falling off.*

Using this method, we can remove the non-informative part of the sentences: *One thing I have to mention is that* and prevent any of the words in that part from being selected as our training features.

Experiment 2 requires the workers to enter or copy and paste text in the box, and 100% of the workers did it. In our sentiment classification experiment described below, we used all of the results without further filtering.

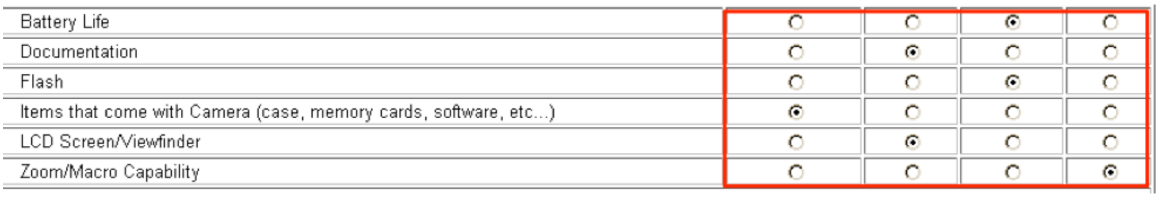
We paid $0.01 for each assignment in both experiments, and we acquired all of the annotations in one week's time with a total cost of $215, including fees paid to Amazon. Our pay rate is about $0.36/hour. For Experiment 1 alone, if we adopted a traditional approach and hired two annotators, they could likely complete the annotations in five 8-hour days. Using this approach, the cost for Experiment 1 alone would be $1200, with a rate of $15/hour. Therefore, our approach is both cheaper and faster than the traditional approach.

Figure 2: Reliable Response



Battery Life	○	●	○	○
Documentation	●	○	○	○
Flash	●	○	○	○
Items that come with Camera (case, memory cards, software, etc...)	●	○	○	○
LCD Screen/Viewfinder	●	○	○	○
Zoom/Macro Capability	●	○	○	○

Figure 3: Unreliable Response



Battery Life	○	○	●	○
Documentation	○	●	○	○
Flash	○	○	●	○
Items that come with Camera (case, memory cards, software, etc...)	●	○	○	○
LCD Screen/Viewfinder	○	●	○	○
Zoom/Macro Capability	○	○	○	●

Having described our crowd-souring based approach for pruning the non-informative features, we next present an empirical study on the effect of training on informative features.

4 Pruning Non-Informative Features for Sentiment Classification

We conducted an experiment on sentiment classification in the domain of camera reviews to test the effect of pruning non-informative features based on AMT workers' annotations.

In our experiment, we select the Nouns, Verbs, Adjectives and Adverbs as our unigram features for training. We define non-informative features as the four types of words occurring in the non-informative portion of the training sentence; namely, the portion that does not mention any aspect of the camera or associated sentiment. For example, for a training sentence such as (1) (repeated below as (6)), training on all features would select the following words: [*film, daughter, ballet, recital, not-believe*[2], *auto, focus, kept, blurring, focusing*].

(6) *I filmed my daughter's ballet recital and could not believe how the auto focus kept blurring then focusing.*

By contrast, pruning non-informative features would yield a shorter list of selected words: [*auto, focus, kept, blurring, focusing*].

In our experiment, we compare the performance of the classifier learned from all of the Nouns, Verbs, Adjectives and Adverbs in the sentences with the one learned from these word types occurring only in the informative part of the sentence. When the training set contains all of the feature words, we refer to it as the All-Features-Set. When the non-informative features are pruned, the training set contains only the informative feature words, which we refer to as the Informative-Features-Set.

All of the feature words are stemmed using the Porter Stemmer (Porter, 1980). Negators are attached to the next selected feature word. We also use a small set of stop words[3] to exclude copulas and words such as *take*. The reason that we choose these words as stop words is because they are both frequent and ambiguous and thus tend to have a negative impact on the classifier.

All of our training and test sentences are annotated through crowd-sourcing techniques as described in the last section. In our experiment we use 2131 sentences in total for training and 587 sentences for hold-out testing. The non-informative part of the test sentences are not removed. The experiment results and implications are discussed in detail in the following subsections.

[2] See below for the description regarding how we handle negation.

[3] The stop words we use include copulas and the following words: *take, takes, make, makes, just, still, even, too, much, enough, back, again, far, same*

4.1 Aspect:Polarity Classification Using SVM

In this experiment, the task is to perform a 45 way sentiment classification. These 45 classes are derived from 22 aspects related to camera purchases such as *picture quality, LCD screen, battery life and customer support* and their associated polarity values *positive* and *negative*, as well as a class of *no opinion* about any of the 22 aspects. An example of such a class is *picture quality: positive*. The classifier maps each input sentence into one of the 45 classes.

One of the approaches we tested is to train the classifier based on the All-Features-Set derived from the original raw sentences. We refer to this as “All Features”. The other approach is to learn from the Informative-Features-Set derived from the sentences with the non-informative portion removed by the AMT workers. We refer to this as “Informative Features”. The experiment is conducted using SVM algorithm implemented by Chang et al. (Chang and Lin, 2001). We use linear kernel type and use the default setting for all other parameters.

The classification accuracy is defined as follows.

$$Accuracy = \frac{NumberofSentencesCorrectlyClassified}{TotalNumberofSentences}. \quad (2)$$

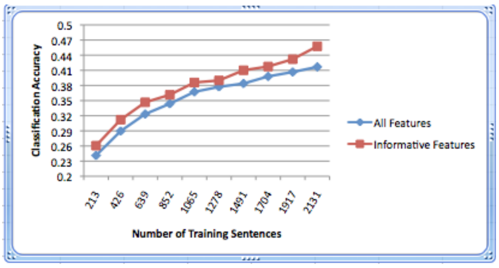
The experiment results in terms of classification accuracy are shown in Table 1.

Table 1: Classification Accuracy

All Features	Informative Features
41.7%	**45.8%**

In this experiment, pruning the non-informative features improves the accuracy by more than 4%. This improvement is statistically significant by a one-tailed sign test at $p = 0.15$. Training on the informative features also consistently improves the classification accuracy when we vary the size of the training data as illustrated by the Figure 5[4].

Figure 5: Learning Curve



A salient characteristic of this experiment is that the training data tend to be very sparse for two reasons. First, the number of classes is large, which means that the number of training examples for each class will be fewer. As shown in Table 2, 24 out of the 45 classes have fewer than 30 training examples, which is an indication of how sparse the training data is. Second, as shown in Section 1, people tend to use different ways to express the type of sentiments that we aim to learn in this experiment. Therefore, it is difficult to collect repeated training examples and this difficulty cannot be simply overcome by increasing the size of the training data. This data sparsity means that it is difficult for the SVM to learn which feature words are non-informative noise.

Table 2: Class Distribution in Experiment 1

Number of Classes	Number of Training Sentences
6	fewer than 10
14	fewer than 20
24	fewer than 30
33	fewer than 50
41	fewer than 100
4	more than 100

4.2 Automatic Feature Selection vs. Pruning by AMT Workers

As shown in the previous subsection, pruning non-informative word features using non-expert annotations can significantly improve the performance of the sentiment classifier. Can we achieve the same improvement by using automatic feature selection algorithms?

We tried three widely used feature selection techniques LR(Likelihood Ratio), WLLR(Weighted Log-Likelihood Ratio) (Nigam et al., 2000; Ng et al., 2006) and MI(Mutual Information) and applied them to the original raw training data. We found that in general, the fewer

[4] To demonstrate the learning curve, we experimented with the use of different percentages of the training sentences while always testing on the same 587 test sentences. When the percentage of the training sentences used is less than 100%, we randomly pick that percentage of training sentences until the test accuracy converges.

the feature words selected by these algorithms, the worse the classifier performs. The classifier performed the best when using all of the available feature words. In other words, automatic feature selection offered no benefit. Table 3 shows the results of using these three automatic feature selection techniques as well as the results of not performing automatic feature selection. The threshold for the LR algorithm was set to be 5; the threshold for the WLLR algorithm was set to be 0.005; and the threshold for the MI algorithm was set to be 2000 (using the top 2000 ranked features out of a total of 3279 features).

Table 3: Automatic Feature Selection Results

No Feature Selection	LR	WLLR	MI
41.7%	35.4%	40.2%	41.1%

This result is not surprising given the data sparsity issue in our experiment. Traditional feature selection methods either try to remove correlated features which can cause havoc for some methods or to prune out features uncorrelated with labels to make learning more efficient. However, we have sparse data so correlations calcuated are very unstable - if a feature appears once with a label what can we conclude? So the same properties that cause difficulties for the learner cause problems for feature selection techniques as well.

To summarize, pruning non-informative word features using non-expert annotations can significantly improve the performance of the sentiment classifier even when the test data still contain non-informative features. We believe this is because pruning non-informative feature words based on human knowledge leads to better training data that cannot be achieved by using automatic feature selection techniques. The subsection below compares the two sets of training sentences we used in this experiment: one comprises the original raw sentences and the other comprises sentences with the non-informative text removed. We show that our approach of pruning non-informative text indeed leads to a better set of training data.

4.3 Comparison of Training Data Before and After the Feature Pruning

Our assumption is that training data is better if data belonging to closer classes are more similar and data belonging to further classes are more different. In our sentiment classification experiment, an example of two very close classes are *battery life: positive* and *battery life: negative*. An example of two very different classes are *battery life: positive* and *auto focus: negative*. The more similar the training data belonging to closer classes and the more dissimilar the training data belonging to different classes, the more accurate the classifier can predict the involved camera aspect, which in turn should lead to improvements on the overall classification accuracy.

To test whether the pruned text produced better training data than the original text, an adjusted cosine similarity measure was used. Note that our measurement can only reflect partial effects of AMT workers' pruning, because our measure is essentially term frequency based, which can reflect similarity in terms of topic (camera aspects in our case) but not similarity in terms of polarity (Pang et al., 2002). Nevertheless, this measurement highlights some of the impact resulting from the pruning.

To compare training data belonging to any two classes, we produce a tf-idf score for each word in those two classes and represent each class as a vector containing the tf-idf score for each word in that class. Comparing the similarity of two classes involves calculating the adjusted cosine similarity in the following formula.

$$similarity = \frac{A \cdot B}{||A||||B||}. \quad (3)$$

A and B in the above formula are vectors of tf-idf scores, whereas in the standard cosine similarity measure A and B would be vectors containing tf scores. The motivation for using tf-idf scores instead of the tf scores is to reduce the importance of highly common words such as *the* and *a* in the comparison. The similarity score produced by this formula is a number between 0 and 1; 0 being no overlap and 1 indicating that the classes are identical. Word stemming was not used in this experiment.

We compared similarity changes in two situations. First, when two classes share the same aspect; this involves comparison between 22 class pairs such as *battery life: positive* vs. *battery life: negative*. Second, when two classes share different aspects; for example, *battery life: positive* vs. *auto focus: negative* and *battery life: positive* vs. *auto focus: positive*. In this situation, we compared the similarity changes in 903 class pairs. If pruning the non-informative text does indeed provide better training data, we expect similarity to increase in the first situation and to decrease in the second situation after the pruning. This is precisely what we found; our finding is summarized in Table 4.

Table 4: Average Similarity Changes in the Pruned Training Data

Same aspect	Different aspect
+0.01	-0.02

In conclusion, AMT workers, by highlighting the most pertinent information for classification and allowing us to discard the rest, provided more useful data than the raw text.

5 Conclusions

To summarize, we found that removing the non-informative text from the training sentences produces better training data and significantly improves the performance of the sentiment classifier even when the test data still contain non-informative feature words. We also show that annotations for both sentiment classes and sentiment-informative texts can be acquired efficiently through crowd-sourcing techniques as described in this paper.

6 Acknowledgments

We thank Prateek Sarkar, Jessica Staddon and Bi Chen for insightful discussions and comments. We thank the anonymous reviewers' helpful suggestions and feedback. We would also like to thank Jason Kessler for implementing part of the sentiment analysis algorithm and the Amazon Mechanical Turk experiments.

References

Sheng, Victor S., Provost, Foster, and Ipeirotis, Panagiotis G.. 2008. *Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers*. KDD 2008 Proceedings 614-622.

Pang, Bo, Lee, Lillian, and Vaithyanathan, Shivakumar. 2002. *Thumbs up? Sentiment classification using machine learning techniques*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 79-86.

Dave, Kushal, Lawrence, Steve, and Pennock, David M.. 2003. *Mining the peanut gallery: Opinion extraction and semantic classification of product reviews*. Proceedings of WWW 519-528.

Kim, Soo–Min and Hovy, Eduard. 2005. *Identifying opinion holders for question answering in opinion texts*. Proceedings of the AAAI Workshop on Question Answering in Restricted Domains.

Wiebe, Janyce M. , Wilson, Theresa , Bruce, Rebecca , Bell, Matthew and Martin, Melanie. 2004. *Learning subjective language*. *Computational Linguistics*, 30(3):277-308.

Yang, Kiduk , Yu, Ning , Valerio, Alejandro and Zhang, Hui. 2006. *WIDIT in TREC-2006 Blog track*. Proceedings of TREC.

Gamon, Michael. 2004. *Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis*. Proceedings of the International Conference on Computational Linguistics (COLING).

Matsumoto, Shotaro, Takamura, Hiroya and Okumura, Manabu. 2005. *Sentiment classification using word sub-sequences and dependency sub-trees*. Proceedings of PAKDD05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining.

Ng, Vincent, Dasgupta, Sajib and Arifin, S. M. Niaz. 2006. *Examining the role of linguistic knowledge sources in the automatic identification and classification of reviews*. Proceedings of the COLING/ACL Main Conference Poster Sessions 611-618.

Mullen, Tony and Collier, Nigel. 2004. *Sentiment analysis using support vector machines with diverse in-formation sources*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 412-418.

Turney, Peter. 2002. *Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews*. Proceedings of the Association for Computational Linguistics (ACL) 417-424.

Whitelaw, Casey, Garg, Navendu and Argamon, Shlomo. 2005. *Using appraisal groups for sentiment analysis*. Proceedings of the ACM SIGIR Conference on Information and Knowledge Management (CIKM) 625-631.

Dillard, Logan. 2007. *I Can't Recommend This Paper Highly Enough: Valence-Shifted Sentences in Sentiment Classification*. Master Thesis.

Riloff, Ellen and Wiebe, Janyce. 2003. *Learning extraction patterns for subjective expressions*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Su, Qi, Pavlov, Dmitry, Chow, Jyh-Herng and Baker, Wendell C.. 2007. *Internet-Scale Collection of Human- Reviewed Data*. Proceedings of WWW-2007.

Nakov, Preslav. 2008. *Paraphrasing Verbs for Noun Compound Interpretation*. Proceedings of the Workshop on Multiword Expressions, LREC-2008.

Kaisser, Michael and Lowe, John B.. 2008. *A Re-search Collection of QuestionAnswer Sentence Pairs*. Proceedings of LREC-2008.

Snow, Rion, O'Connor, Brendan, Jurafsky, Daniel and Ng, Andrew Y. 2008. *Cheap and Fast - But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks*. Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP).

Blitzer, John, Dredze, Mark, Biographies, Fernando Pereira., Bollywood, Boom-boxes and Blenders. 2007. *Domain Adaptation for Sentiment Classification*. Proceedings of the Association for Computational Linguistics (ACL).

Porter, M.F.. 1980. *An algorithm for suffix stripping*. Program.

Chang, Chih-Chung and Lin, Chih-Jen. 2001. *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Nigam, K., McCallum, A.K., Thrun, S., and Mitchell, T.. 2000. *Text Classification from labeled and unlabeled documents using em*. Machine Learning 39(2-3) 103-134.

Measuring Conceptual Similarity by Spreading Activation over Wikipedia's Hyperlink Structure

Stephan Gouws, G-J van Rooyen, and Herman A. Engelbrecht
Stellenbosch University
Stellenbosch, South Africa.
{stephan,gvrooyen,hebrecht}@ml.sun.ac.za

Abstract

Keyword-matching systems based on simple models of semantic relatedness are inadequate at modelling the ambiguities in natural language text, and cannot reliably address the increasingly complex information needs of users. In this paper we propose novel methods for computing semantic relatedness by spreading activation energy over the hyperlink structure of Wikipedia. We demonstrate that our techniques can approach state-of-the-art performance, while requiring only a fraction of the background data.

1 Introduction

The volume of information available to users on the World Wide Web is growing at an exponential rate (Lyman and Varian, 2003). Current keyword-matching information retrieval (IR) systems suffer from several limitations, most notably an inability to accurately model the ambiguities in natural language, such as synonymy (different words having the same meaning) and polysemy (one word having multiple different meanings), which is largely governed by the context in which a word appears (Metzler and Croft, 2006).

In recent years, much research attention has therefore been given to *semantic* techniques of information retrieval. Such systems allow for sophisticated semantic search, however, require the use of a more difficult-to-understand query-syntax (Tran et al., 2008). Furthermore, these methods require specially encoded (and thus costly) *ontologies* to describe the particular domain knowledge in which the system operates, and the specific interrelations of concepts within that domain.

In this paper, we focus on the problem of computationally estimating similarity or relatedness between two natural-language documents. A novel technique is proposed for computing semantic similarity by spreading activation over the hyperlink structure of Wikipedia, the largest free online encyclopaedia. New measures for computing similarity between individual concepts (**inter-concept similarity**, such as "France" and "Great Britain"), as well as between documents (**inter-document similarity**) are proposed and tested. It will be demonstrated that the proposed techniques can achieve comparable inter-concept and inter-document similarity accuracy on similar datasets as compared to the current state of the art Wikipedia Link-based Measure (WLM) (Witten and Milne, 2008) and Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007) methods respectively. Our methods outperform WLM in computing inter-concept similarity, and match ESA for inter-document similarity. Furthermore, we use the same background data as for WLM, which is less than 10% of the data required for ESA.

In the following sections we introduce work related to our work and an overview of our approach and the problems that have to be solved. We then discuss our method in detail and present several experiments to test and compare it against other state-of-the-art methods.

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 46–54,

2 Related Work and Overview

Although Spreading Activation (SA) is foremost a cognitive theory modelling semantic memory (Collins and Loftus, 1975), it has been applied computationally to IR with various levels of success (Preece, 1982), with the biggest hurdle in this regard the cost of creating an associative network or knowledge base with adequate conceptual coverage (Crestani, 1997). Recent knowledge-based methods for computing semantic similarity between texts based on Wikipedia, such as Wikipedia Link-based Measure (WLM) (Witten and Milne, 2008) and Explicit Semantic Analysis (ESA) (Gabrilovich and Markovitch, 2007), have been found to outperform earlier WordNet-based methods (Budanitsky and Hirst, 2001), arguably due to Wikipedia's larger conceptual coverage.

WLM treats the anchor text in Wikipedia articles as links to other articles (all links are treated equally), and compare concepts based on how much overlap exists in the out-links of the articles representing them. ESA discards the link structure and uses only the text in articles to derive an explicit concept space in which each dimension represents one article/concept. Text is categorised as vectors in this concept space and similarity is computed as the cosine similarity of their ESA vectors. The most similar work to ours is Yeh (2009) in which the authors derive a graph structure from the inter-article links in Wikipedia pages, and then perform random walks over the graph to compute relatedness.

In Wikipedia, users create links between articles which are seen to be related to some degree. Since links relate one article to its neighbours, and by extension to their neighbours, we extract and process this hyperlink structure (using SA) as an **Associative Network (AN)** (Berger et al., 2004) of concepts and links relating them to one another. The SA algorithm can briefly be described as an iterative process of propagating real-valued energy from one or more source nodes, via weighted links over an associative network (each such a propagation is called a *pulse*). The algorithm consists of two steps: First, one or more pulses are triggered, and second, termination checks determine whether the process should continue or halt. This process of activating more and more nodes in the network and checking for termination conditions are repeated pulse after pulse, until all termination conditions are met, which results in a final activation state for the network. These final node activations are then translated into a score of relatedness between the initial nodes.

Our work presents a computational implementation of SA over the Wikipedia graph. We therefore overcome the cost of producing a knowledge base of adequate coverage by utilising the collaboratively-created knowledge source Wikipedia. However, additional strategies are required for translating the hyperlink structure of Wikipedia into a suitable associative network format, and for this new techniques are proposed and tested.

3 Extracting the Hyperlink Graph Structure

One article in Wikipedia covers one specific topic (concept) in detail. Hyperlinks link a page A to a page B, and are thus *directed*. We can model Wikipedia's hyperlink structure using standard graph theory as a *directed graph* G, consisting of a set of vertices $\mathbf{V}$, and a set of edges $\mathbf{E}$. Each edge $e_{ij} \in \mathbf{E}$ connects two vertices $v_i, v_j \in \mathbf{V}$. For consistency, we use the term *node* to refer to a vertex (Wikipedia article) in the graph, and *link* to refer to an edge (hyperlink) between such nodes.

In this model, each Wikipedia article is seen to represent a single *concept*, and the hyperlink structure relates these concepts to one another. In order to compute relatedness between two concepts v_i and v_j, we use spreading activation and rely on the fundamental principle of an associative network, namely that it connects nodes that are associated with one another via real-valued links denoting how strongly the objects are related. Since Wikipedia was not created as an associative network, but primarily as an online encyclopaedia, none of these weights exist, and we will have to deduce these (see *Fan-out constraint* in Section 4).

Links *into* pages are used, since this leads to better results (Witten and Milne, 2008). The Wikipedia graph structure is represented in an adjacency list structure, i.e. for each node v_i we store its list of neighbour nodes in a dictionary using v_i's id as key. This approach is preferred over an adjacency matrix structure, since most articles are linked to by only 34 articles on average, which would lead to a very sparse adjacency matrix structure.

4 Adapting Spreading Activation for Wikipedia's Hyperlink Structure

Each pulse in the Spreading Activation (SA) process consists of three stages: 1) pre-adjustment, 2) spreading, and 3) post-adjustment (Crestani, 1997). During pre- and post-adjustment, some form of activation decay is optionally applied to the active nodes. This serves both to avoid retention of activation from previous pulses, and, from a connectionist point of view, models 'loss of interest' when nodes are not continually activated.

Let $a_{i,\text{in}}$ denote the total energy input (activation) for node v_i, and $N(v_i)$ the set of v_i's neighbour nodes with incoming links to v_i. Also, let $a_{j,\text{out}}$ denote the output activation of a node v_j connected to node v_i, and let w_{ij} denote the weight of connection between node v_i and v_j. For a node v_i, we can then describe the pure model of spreading activation as follows:

$$a_{i,\text{in}} = \sum_{v_j \in N(v_i)} a_{j,\text{out}} w_{ij}. \qquad (1)$$

This pure model of SA has several significant problems, the most notable being that activation can saturate the entire network unless certain constraints are imposed, namely limiting how far activation can spread from the initially activated nodes (distance constraint), and limiting the effect of very highly-connected nodes (fan-out constraint) (Crestani, 1997). In the following three sections we discuss how these constraints were implemented in our model for SA.

Distance constraint

For every pulse in the spreading process, a node's activation value is multiplied by a global network decay parameter $0 < d < 1$. We therefore substitute w_{ij} in Equation 1 for $w_{ij}d$. This decays activation exponentially in the path length. For a path length of one, activation is decayed by d, for a path length of two, activation is decays by $dd = d^2$, etc. This penalises activation transfer over longer paths. We also include a maximum path length parameter $L_{p,\max}$ which limits how far activation can spread.

Fan-out constraint

As noted above, in an associative network, links have associated real-valued weights to denote the strength of association between the two nodes they connect (i.e. w_{ij} in Equation 1). These weights have to be estimated for the Wikipedia hyperlink graph, and for this purpose we propose the use of three **weighting schemes**:

In **pure Energy Distribution (ED)**, a node v_i's weight w is made inversely proportional to its in-degree (number of neighbours $N(v_i) \geq 1$ with incoming links to v_i[1]). Thus $\text{ED}(v_i, v_j) = w_{ij} = \frac{1}{|N(v_i)|}$. This reduces the effect of very connected nodes on the spreading process (constraint 2 above).

For instance, we consider a path connecting two nodes via a general article such as **`USA`** (connected to 322,000 articles) not nearly as indicative of a semantic relationship, as a path connecting them via a very **specific** concept, such as **`Hair Pin`** (only connected to 20 articles).

Inverse Link-Frequency (ILF) is inspired by the term-frequency inverse document-frequency (**tf-idf**) heuristic (Salton and McGill, 1983) in which a term's weight is reduced as it is contained in more documents in the corpus. It is based on the idea that the more a term appears in documents across the corpus, the less it can discriminate any one of those documents.

We define a node v_i's *link-frequency* as the number of nodes that v_i is connected to $|N(v_i)|$ divided by the number of possible nodes it could be connected to in the entire Wikipedia graph

[1] All orphan nodes are removed from the AN.

$|G|$, and therefore give the log-smoothed *inverse* link-frequency of node v_i as:

$$\mathrm{ILF}(v_i) \triangleq \log\left(\frac{|G|}{|N(v_i)|}\right) \geq 0 \qquad (2)$$

As noted above for *pure energy distribution*, we consider less connected nodes as more specific. If one node connects to another via a very **specific** node with a low in-degree, $\frac{|G|}{|N(v_i)|}$ is very large and $\mathrm{ILF}(v_i) > 1$, thus boosting that specific link's weight. This has the effect of 'boosting' paths (increasing their contribution) which contain nodes that are less connected, and therefore more meaningful in our model.

To evaluate the effect of this boosting effect described above, we also define a third normalised weighting scheme called the **Normalised Inverse Link-Frequency (NILF)**, $0 \leq \mathrm{NILF}(v_i) \leq 1$:

$$\mathrm{NILF}(v_i) \triangleq \frac{\mathrm{ILF}(v_i)}{\log|G|}. \qquad (3)$$

ILF reaches a maximum of $\log|G|$ when $|N(v_i)| = 1$ (see Equation 2). We therefore divide by $\log|G|$ to normalise its range to [0,1].

Threshold constraint

Finally, the above-mentioned constraints are enforced through the use of a threshold parameter $0 < T < 1$. Activation transfer to a next node ceases when a node's activation value drops below a certain threshold T.

5 Strategies for Interpreting Activations

After spreading has ceased, we are left with a vector of nodes and their respective values of activation (an **activation vector**). We wish to translate this activation vector into a score resembling strength of association or relatedness between the two initial nodes.

We approach this problem using two different approaches, the Target Activation Approach (TAA) and the Agglomerative Approach (AA). These approaches are based on two distinct hypotheses, namely: Relatedness between two nodes can be measured as either 1) the ratio of initial energy that reaches the target node, or 2) the amount of overlap between their individual activation vectors by spreading from both nodes individually.

Target Activation Approach (TAA)

To measure the relatedness between v_i and v_j, we set a_i to some initial value K_{init} (usually 1.0), and all node activations including $a_j = 0$. After the SA process has terminated, v_j is activated with some $a_{j,\mathrm{in}}$. Relatedness is computed as the ratio $\mathrm{sim}_{TAA}(v_i, v_j) \triangleq \frac{a_{j,\mathrm{in}}}{K_{\mathrm{init}}}$.

Agglomerative Approach (AA)

The second approach is called the Agglomerative Approach since we agglomerate all activations into one score resembling relatedness. After spreading has terminated, relatedness is computed as the amount of overlap between the individual nodes' activation vectors, using either the cosine similarity (AA-cos), or an adapted version of the information theory based WLM (Witten and Milne, 2008) measure.

Assume the same set of initial nodes v_i and v_j. Let $\mathbf{A}_k$ be the N-dimensional vector of real-valued activation values obtained by spreading over the N nodes in the graph from node v_k (called an **activation vector**). We use a_{kx} to denote the element at position x in $\mathbf{A}_k$. Furthermore, let $\mathbf{V}_k = \{v_{k1}, ..., v_{kM}\}$ denote the set of M nodes activated by spreading from v_k, i.e. the set of identifiers of nodes with non-zero activations in $\mathbf{A}_k$ after spreading has terminated (and therefore $M \leq N$).

We then define the **cosine Agglomerative Approach** (henceforth called **AA-cos**) as

$$\begin{aligned} &\mathrm{sim}_{\mathrm{AA,cos}}(\mathbf{A}_i, \mathbf{A}_j) \\ &\triangleq \frac{\mathbf{A}_i \cdot \mathbf{A}_j}{||\mathbf{A}_i||||\mathbf{A}_j||} \end{aligned} \qquad (4)$$

For our adaptation of the Wikipedia Link-based Measure (WLM) approach to spreading activation, we define the **WLM Agglomerative Approach** (henceforth called **AA-wlm**[2]) as

[2] *AA-wlm* is our adaptation of WLM (Witten and Milne, 2008) for SA, not to be confused with their method, which we simply call *WLM*.

$$\mathrm{sim}_{\mathrm{AA,wlm}}(\mathbf{V_i}, \mathbf{V_j}) \triangleq \frac{\log\big(\max(|\mathbf{V_i}|,|\mathbf{V_j}|)\big)-\log(|\mathbf{V_i}\cap\mathbf{V_j}|)}{\log(|G|)-\log(\min(|\mathbf{V_i}|,|\mathbf{V_j}|))} \quad (5)$$

with $|G|$ representing the number of nodes in the entire Wikipedia hyperlink graph. Note that the AA-wlm method does not take activations into account, while the AA-cos method does.

6 Spreading Activation Algorithm

Both the TAA and AA approaches described above rely on a function to spread activation from one node to all its neighbours, and iteratively to all their neighbours, subject to the constraints listed. TAA stops at this point and computes relatedness as the ratio of energy received to energy sent between the target and source node respectively. However, AA repeats the process from the target node and computes relatedness as some function (cosine or information theory based) of the two activation vectors, as given by Equation 4 and Equation 5.

We therefore define SPREAD_UNIDIR() as shown in Algorithm 1. Prior to spreading from some node v_i, its activation value a_i is set to some initial activation value K_{init} (usually 1.0). The activation vector $\mathbf{A}$ is a dynamic node-value-pair list, updated in-place. $\mathbf{P}$ is a dynamic list of nodes in the *path* to v_i to avoid cycles.

Algorithm 1 Pseudo code to spread activation depth-first from node v_i up to level $L_{p,\max}$, using global decay d, and threshold T, given an adjacency list graph structure G and a weighting scheme $\mathbf{W}$ such that $0 < w_{ij} \in \mathbf{W} < 1$.

```
Require: $G, L_{p,\max}, d, T$
  function SPREAD_UNIDIR($v_i, \mathbf{A}, \mathbf{P}$)
    if $(v_i, a_i) \notin \mathbf{A}$ or $a_i < T$ then        ▷ Threshold
      return
    end if
    Add $v_i$ to $\mathbf{P}$                                  ▷ To avoid cycles
    for $v_j \in N(v_i)$ do                                    ▷ Process neighbours
      if $(v_j, a_j) \notin \mathbf{A}$ then
        $a_j = 0$
      end if
      if $v_j \notin \mathbf{P}$ and $|\mathbf{P}| \leq L_{p,\max}$ then
        $a_j^* = a_j + a_i * w_{ij} * d$
        Replace $(v_j, a_j) \in \mathbf{A}$ with $(v_j, a_j^*)$
        SPREAD_UNIDIR($v_j, \mathbf{A}, \mathbf{P}$)
      end if
    end for
    return
  end function
```

7 Parameter Optimisation: Inter-concept Similarity

The model for SA as introduced in this paper relies on several important parameters, namely the spreading strategy (TAA, AA-cos, or AA-wlm), weighting scheme (pure ED, ILF, and NILF), maximum path length $L_{p,\max}$, network decay d, and threshold T. These parameters have a large influence on the accuracy of the proposed technique, and therefore need to be optimised.

Experimental Method

In order to compare our method with results reported by Gabrilovich and Markovitch (2007) and Witten and Milne (2008), we followed the same approach by randomly selecting 50 word-pairs from the WordSimilarity-353 dataset (Gabrilovich, 2002) and correlating our method's scores with the human-assigned scores. To reduce the possibility of overestimating the performance of our technique on a sample set that happens to be favourable to our technique, we furthermore implemented a technique of **repeated holdout** (Witten and Frank, 2005):

Given a sample test set of N pairs of words with human-assigned ratings of relatedness, randomly divide this set into k parts of roughly equal size[3]. Hold out one part of the data and iteratively evaluate the performance of the algorithm on the remaining $k-1$ parts until all k parts have been held out once. Finally, average the algorithm's performance over all k runs into one score resembling the performance for that set of parameters.

Since there are five parameters (spreading strategy, weighting scheme, path length, network decay, and threshold), a **grid search** was implemented by holding three of the five parameters constant, and evaluating combinations of decay and threshold by stepping over the possible parameter space using some step size. A coarse-grained grid search was first conducted with step

[3] k was chosen as 5.

Table 1: Spreading results by spreading strategy (TAA=Target Activation Approach, AA=Agglomerative Approach, $L_{p,\max}$ = maximum path length used, ED=energy distribution only, ILF=Inverse Link Frequency, NILF=normalised ILF.) Best results in bold.

Strategy	ρ_{max}	Parameters
TAA	0.56	ED, $L_{p,\max}$=3, d=0.6, T=0.001
AA-wlm	0.60	NILF, $L_{p,\max}$=3, d=0.1, T=10^{-6}
AA-cos	0.70	ILF, $L_{p,\max}$=3, d=0.5, T=0.1

size of 0.1 over d and a logarithmic scale over T, thus $T = \{0, 0.1, 0.01, 0.001, ..., 10^{-9}\}$. The best values for d and T were then chosen to conduct a finer-grained grid search.

Influence of the different Parameters

The **spreading strategy** determines how activations resulting from the spreading process are converted into scores of relatedness or similarity between two nodes. Table 1 summarises the best results obtained for each of the three strategies, with the specific set of parameters that were used in each run.

Results are better using the AA (ρ_{max} = 0.70 for AA-cos) than using the TAA (ρ_{max} = 0.56). Secondly, the AA-cos spreading strategy significantly outperforms the AA-wlm strategy over this sample set ($\rho_{\max,\text{wlm}}$ = 0.60 vs $\rho_{\max,\cos}$ = 0.70). These results compare favourably to similar inter-concept results reported for WLM (Witten and Milne, 2008) ($\rho = 0.69$) and ESA (Gabrilovich and Markovitch, 2007) ($\rho = 0.75$).

Maximum path length $L_{p,\max}$ is related to how far one node can spread its activation in the network. We extend the first-order link model used by WLM, by approaching the link structure as an associative network and by using spreading activation.

To evaluate if this is a useful approach, tests were conducted by using maximum path lengths of one, two, and three. Table 2 summarises the results for this experiment. Increasing path length from one to two hops increases performance from $\rho_{max} = 0.47$ to $\rho_{max} = 0.66$. Moreover, increasing $L_{p,\max}$ from two to three hops furthermore increases performance to $\rho_{max} = 0.70$.

Table 2: Spreading results by maximum path length $L_{p,\max}$. Best results in bold.

$L_{p,\max}$	ρ_{max}	Parameters
1	0.47	TAA, ED/ILF/NILF
2	0.66	AA-cos, ILF, d=0.4, T=0.1
3	0.70	AA-cos, ILF, d=0.5, T=0.1

Table 3: Spreading results by weighting scheme w. Best results in bold.

w	ρ_{max}	Parameters
NILF	0.63	AA-cos, $L_{p,\max}$ = 3, d=0.9, T=0.01
ED	0.64	AA-cos, $L_{p,\max}$ = 3, d=0.9, T=0.01
ILF	0.70	AA-cos, $L_{p,\max}$ = 3, d=0.5, T=0.1

In an associative network, each link has a real-valued weight denoting the *strength of association* between the two nodes it connects. The derived Wikipedia hyperlink graph lacks these weights. We therefore proposed three new **weighting schemes** (pure ED, ILF, and NILF) to estimate these weights.

Table 3 summarises the best performances using the different weighting schemes. ILF outperforms both ED and NILF. Furthermore, both ED and NILF perform best using higher decay values (both 0.9) and lower threshold values (both 0.01), compared to ILF (0.5 and 0.1 respectively for d and T). We attribute this observation to the boosting effect of the ILF weighting scheme for less connected nodes, and offer the following explanation:

Recall from the section on ILF that in our model, strongly connected nodes are viewed as more general, and nodes with low in-degrees are seen as very **specific** concepts. We argued that a path connecting two concepts via these more specific concepts are more indicative of a stronger semantic relationship than through some very general concept. In the ILF weighting scheme, paths containing these less connected nodes are automatically boosted to be more important. Therefore, by not boosting less meaningful paths, a lower decay and higher threshold effectively limits the amount of non-important nodes that are activated, since their activations are more quickly decayed, whilst at the same time requiring a higher threshold to continue spreading. Boosting more important nodes can therefore lead to activation vectors which capture the semantic context of the source nodes more accurately, leading to higher performance.

8 Computing document similarity

To compute document similarity, we first extract key representative Wikipedia concepts from a document to produce **document concept vectors**[4]. This process is known as *wikification* (Csomai and Mihalcea, 2008), and we used an implementation of Milne and Witten (2008). This produces document concept vectors of the form $\mathbf{V_i} = \{(id_1, w_1), (id_2, w_2), ...\}$ with id_i some Wikipedia article identifier and w_i a weight denoting how strongly the concept relates to the current document. We next present two algorithms, MAXSIM and WIKISPREAD, for computing document similarity, and test these over the Lee (2005) document similarity dataset, a set of 50 documents between 51 and 126 words each, with the averaged gold standard similarity ratings produced by 83 test subjects (see (Lee et al., 2005)).

The first metric we propose is called **MAXSIM** (see Algorithm 2) and is based on the idea of measuring document similarity by pairing up each Wikipedia concept in one document's concept vector with its most similar concept in the other document. We average those similarities to produce an inter-document similarity score, weighted by how strongly each concept is seen to represent a document ($0 < p_i < 1$). The contribution of a concept is further weighted by its ILF score, so that more specific concepts contribute more to final relatedness.

The second document similarity metric we propose is called the **WIKISPREAD** method and is a natural extension of the inter-concept spreading activation work introduced in the previous section. We view a document concept vector as a cluster of concepts, and build a single *document activation vector* (see Algorithm 3) – i.e. a vector of article ids and their respective activations – for each document, by iteratively spreading from each concept in the document concept vector. Finally, similarity is computed using either the AA-cos or AA-wlm methods given by Equation 4 and Equation 5 respectively.

Algorithm 2 Pseudo code for the MaxSim algorithm for computing inter-document similarity. v_i is a Wikipedia concept and $0 < p_i < 1$ how strongly it relates to the current document.

```
Require: ILF lookup function
  function MAXSIM($\mathbf{V_1}, \mathbf{V_2}$)
      num=0
      den=0
      for $(v_i, p_i) \in \mathbf{V_1}$ do
          $s_k = 0$                          ▷ $s_k = \max_j \mathrm{sim}(v_i, v_j)$
          for $v_j \in \mathbf{V_2}$ do              ▷ Find most related topic
              $s_j = \mathrm{sim}(v_i, v_j)$
              if $s_j > s_k$ then
                  $v_k = v_j$ ▷ Topic in $\mathbf{V_2}$ most related to $v_i$
                  $s_k = s_j$
              end if
          end for
          num += $s_k p_i \mathrm{ILF}(v_k)$
          den += $\mathrm{ILF}(v_k)$
      end for
      return num / den
  end function
```

Algorithm 3 Pseudo code for the WikiSpread algorithm for computing inter-document similarity. $K_{\text{init}} = 1.0$.

```
  function WIKISPREAD($\mathbf{V_1}, \mathbf{V_2}$)
      $\mathbf{A_1} = \emptyset$                  ▷ Dynamic activation vectors.
      $\mathbf{A_2} = \emptyset$
      for $(v_i, p_i) \in \mathbf{V_1}$ do                ▷ Document 1
          $a_i = K_{\text{init}} \cdot p_i$              ▷ Update $a_i \propto p_i$
          Add $(v_i, a_i)$ to $\mathbf{A_1}$
          SPREAD_UNIDIR($v_i, \mathbf{A_1}, \emptyset$)
      end for
      for $(v_j, p_j) \in \mathbf{V_2}$ do                ▷ Document 2
          $a_j = K_{\text{init}} \cdot p_j$
          Add $(v_j, a_j)$ to $\mathbf{A_2}$
          SPREAD_UNIDIR($v_j, \mathbf{A_2}, \emptyset$)
      end for
      Compute similarity using AA-cos or AA-wlm
  end function
```

Knowledge-based approaches such as the Wikipedia-based methods can capture more complex lexical and semantic relationships than

[4] Vectors of Wikipedia topics (concepts) and how strongly they are seen to relate to the current document.

Table 4: Summary of final document similarity correlations over the Lee & Pincombe document similarity dataset. ESA score from Gabrilovich and Markovitch (2007).

	Pearson ρ
Cosine VSM (with tf-idf) only	0.56
MaxSim method	0.68
WikiSpread method	0.62
ESA	0.72
Combined (Cosine + MaxSim)	**0.72**

keyword-matching approaches, however, nothing can be said about concepts not adequately represented in the underlying knowledge base (Wikipedia). We therefore hypothesise that combining the two approaches will lead to more robust document similarity performance. Therefore, the final document similarity metric we evaluate (**COMBINED**) is a linear combination of the best-performing Wikipedia-based methods described above, and the well-known Vector Space Model (VSM) with cosine similarity and tf-idf (Salton and McGill, 1983).

Results

The results obtained on the Lee (2005) document similarity dataset using the three document similarity metrics (MAXSIM, WIKISPREAD, and COMBINED) are summarised in Table 4. Of the two Wikipedia-only methods, the MaxSim method achieves the best correlation score of $\rho = 0.68$. By combining the standard cosine VSM with tf-idf with the MaxSim metric in the ratio λ and $(1 - \lambda)$ for $0 < \lambda < 1$, and performing a parameter sweep over λ, we can weight the contributions made by the individual methods and observe the effect this has on final performance. The results are shown in Fig 1. Note that both methods contribute equally ($\lambda = 0.5$) to the final best correlation score of $\rho = 0.72$. This suggests that selective knowledge-based augmentation of simple VSM methods can lead to more accurate document similarity performance.

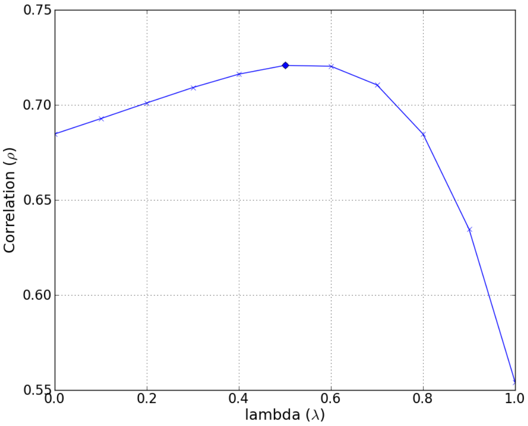


Figure 1: Parameter sweep over λ showing contributions from cosine (λ) and Wikipedia-based MAXSIM method $(1 - \lambda)$ to the final performance over the Lee (2005) dataset.

9 Conclusion

In this paper, the problem of computing conceptual similarity between concepts and documents are approached by spreading activation over Wikipedia's hyperlink graph. New strategies are required to infer weights of association between articles, and for this we introduce and test three new weighting schemes and find our Inverse Link-Frequency (ILF) to give best results. Strategies are also required for translating resulting activations into scores of relatedness, and for this we propose and test three new strategies, and find that our cosine Agglomerative Approach gives best results. For computing document similarity, we propose and test two new methods using only Wikipedia. Finally, we show that using our best Wikipedia-based method to augment the cosine VSM method using tf-idf, leads to the best results. The final result of $\rho = 0.72$ is equal to that reported for ESA (Gabrilovich and Markovitch, 2007), while requiring less than 10% of the Wikipedia database required for ESA. Table 4 summarises the document-similarity results.

Acknowledgements

We thank Michael D. Lee for his document similarity data and MIH Holdings Ltd. for financially supporting this research.

References

Berger, Helmut, Michael Dittenbach, and Dieter Merkl. 2004. An adaptive information retrieval system based on associative networks. *APCCM '04: Proceedings of the first Asian-Pacific conference on Conceptual Modelling*, pages 27–36.

Budanitsky, A. and G. Hirst. 2001. Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In *Workshop on WordNet and Other Lexical Resources*, volume 2. Citeseer.

Collins, A.M. and E.F. Loftus. 1975. A spreading-activation theory of semantic processing. *Psychological review*, 82(6):407–428.

Crestani, F. 1997. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review*, 11(6):453–482.

Csomai, A. and R. Mihalcea. 2008. Linking documents to encyclopedic knowledge. *IEEE Intelligent Systems*, 23(5):34–41.

Gabrilovich, E. and S. Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 6–12.

Gabrilovich, E. 2002. The WordSimilarity-353 Test Collection. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*.

Lee, M.D., B. Pincombe, and M. Welsh. 2005. A Comparison of Machine Measures of Text Document Similarity with Human Judgments. In *27th Annual Meeting of the Cognitive Science Society (CogSci2005)*, pages 1254–1259.

Lyman, P. and H.R. Varian. 2003. How much information? `http://www2.sims.berkeley.edu/research/projects/how-much-info-2003/index.htm`. Accessed: May, 2010.

Metzler, Donald and W. Bruce Croft. 2006. Beyond bags of words: Modeling implicit user preferences in information retrieval. *AAAI'06: Proceedings of the 21st National Conference on Artificial Intelligence*, pages 1646–1649.

Milne, David and Ian H. Witten. 2008. Learning to link with wikipedia. *CIKM '08: Proceeding of the 17th ACM Conference on Information and Knowledge Management*, pages 509–518.

Preece, SE. 1982. *Spreading Activation Network Model for Information Retrieval.* Ph.D. thesis.

Salton, G. and M.J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill New York.

Tran, T., P. Cimiano, S. Rudolph, and R. Studer. 2008. Ontology-based Interpretation of Keywords for Semantic Search. *The Semantic Web*, pages 523–536.

Witten, I.H. and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann.

Witten, I.H. and D. Milne. 2008. An Effective, Low-Cost Measure of Semantic Relatedness Obtained From Wikipedia Links. In *Proceeding of AAAI Workshop on Wikipedia and Artificial Intelligence: an Evolving Synergy, AAAI Press, Chicago, USA*, pages 25–30.

Yeh, E., D. Ramage, C.D. Manning, E. Agirre, and A. Soroa. 2009. WikiWalk: Random walks on Wikipedia for semantic relatedness. In *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing*, pages 41–49. Association for Computational Linguistics.

Identifying and Ranking Topic Clusters in the Blogosphere

M. Atif Qureshi
Korea Advanced Institute of
Science and Technology
atifms@kaist.ac.kr

Arjumand Younus
Korea Advanced Institute of
Science and Technology
arjumandms@kaist.ac.kr

Muhammad Saeed
University of Karachi
saeed@uok.edu.pk

Nasir Touheed
Institute of Business Administration
ntouheed@iba.edu.pk

Abstract

The blogosphere is a huge collaboratively constructed resource containing diverse and rich information. This diversity and richness presents a significant research challenge to the Information Retrieval community. This paper addresses this challenge by proposing a method for identification of "topic clusters" within the blogosphere where topic clusters represent the concept of grouping together blogs sharing a common interest i.e. topic, the algorithm takes into account both the hyperlinked social network of blogs along with the content in the blog posts. Additionally we use various forms and parts-of-speech of the topic to provide a broader coverage of the blogosphere. The next step of the method is to assign topic-specific ranks to each blog in the cluster using a metric called "Topic Discussion Rank," that helps in identifying the most influential blog for a specific topic. We also perform an experimental evaluation of our method on real blog data and show that the proposed method reaches a high level of accuracy.

1 Introduction

With a proliferation of Web 2.0 services and applications there has been a major paradigm shift in the way we envision the World Wide Web (Anderson, 2007; O'Reilly, 2005). Previously the Web was considered as a medium to access information in a read-only fashion. Weblogs or blogs is one such application that has played an effective role in making the Web a social gathering point for masses. The most appealing aspect of blogs is the empowerment they provide to people on the World Wide Web by enabling them to publish their own opinions, ideas, and thoughts on many diverse topics of their own interest generally falling into politics, economics, sports, technology etc. A blog is usually like a personal diary (Sorapure, 2003) with the difference that it's now online and accessible to remote people, it consists of posts arranged chronologically by date and it can be updated on a regular basis by the author of the blog known as blogger. Moreover bloggers have the option to link to other blogs thereby creating a social network within the world of blogs called the blogosphere – in short the blogosphere is a collaboratively constructed resource with rich information on a wide spectrum of topics having characteristics very different from the traditional Web.

However with these differing characteristics of blogs arise many research challenges and this is in particular the case for the Information Retrieval domain. One important problem that arises within this huge blogosphere (Sifry, 2009) is with respect to identification of topic clusters. Such a task involves identification of the key blog clusters that share a common interest point (i.e., topic) reflected quite frequently through

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 55–62,

their blog posts. This is a special type of clustering problem with useful applications in the domain of blog search as Mishne and de Rijke (2006) point out in their study of blog search about the *concept queries* submitted by users of blog search systems.

Moreover ranking these bloggers with respect to their interest in the topic is also a crucial task in order to recognize the most influential blogger for that specific topic. However the blog ranking problem has a completely different nature than the web page ranking problem and link popularity based algorithms cannot be applied for ranking blogs. The reasons for why link based methods cannot be used for blog ranking are as follows:

- Blogs have very few links when compared to web pages; Leskovec et al. report that average number of links per blog post is only 1.6 links (2007). This small number of links per blog results in formation of very sparse network especially when trying to find blogs relevant to a particular topic.
- Blog posts are associated with a timestamp and they need some time for getting in-links. In most of the cases when they receive the links the topics which they talk about die out.
- When link based ranking techniques are used for blogs, bloggers at times assume the role of spammers and try to exploit the system to boost rank of their blogs.

In this paper we propose a solution for identification of topic clusters from within the blogosphere for any topic of interest. We also devise a way to assign topic-specific ranks for each identified blog within the topic cluster. The cluster is identified by the calculation of a metric called “Topic Discussion Isolation Rank (TDIR).” Each blog in the cluster is also assigned a topic rank by further calculation of another metric “Topic Discussion Rank (TDR).” The first metric "TDIR" is applied to a blog in isolation for the topic under consideration and the second metric "TDR" takes into account the blog’s role in its neighborhood for that specific topic. Our work differs from past approaches (Kumar et al., 2003; Gruhl et al., 2004; Chi et al., 2007; Li et al., 2009) in that it takes into consideration both the links between the blogs as well as the content in the blog posts whereas a majority of the past methods follow only link structure. Furthermore we make use of some natural language processing techniques to ensure better coverage of our cluster-finding and ranking methodology. We also perform an experimental evaluation of our proposed solution and release the resultant data of blog clusters and the ranks as an XML corpus.

The remainder of this paper is organized as follows. Section 2 presents a brief summary of related work in this dimension and explains how our proposed methodology differs from these works. Section 3 explains the concept of “topic clusters” in detail along with a description of our solution for clustering and ranking blogs on basis of topics. Section 4 explains our experimental methodology and presents our experimental evaluations on a corpus of 50,471 blog posts gathered from 102 blogs. Section 5 concludes the paper with a discussion of future work in this direction.

2 Related Work

Given the vast amount of useful information in the blogosphere there have been many research efforts for mining and analysis of the blogosphere. This section reviews some of the works that are relevant to our study.

There have been several works with respect to community detection in the blogosphere: one of the oldest works in this dimension is by Kumar et al. who studied the bursty nature of the blogosphere by extracting communities using the hyperlinks between the blogs (2003). Gruhl et al. proposed a transmission graph to study the flow of information in the blogosphere and the proposed model is based on disease-propagation model in epidemic studies (2004). Chi et al. studied the evolution of blog communities over time and introduced the concept of *community factorization* (2007). A fairly recent work is by Li et al. that studies the information propagation pattern in the blogosphere through *cascade affinity* which is an inclination of a blogger to join a particular blog community (2009). Apart from detection of communities within the blogosphere another related study which has recently attracted much interest is of identifying influentials within a "blog community" (Nakajima et al., 2005; Agarwal et al., 2008). All these works base their analysis on link structure of the blogosphere whereas our analytical model differs from these works in that it assigns topic based ranks to the blogs by taking into account both links and blog post's contents.

Along with the community detection problem in the blogosphere there has also been an increasing interest in ranking blogs. Fujimura et al. point out the weak nature of hyperlinks in the web blogs and due to that nature they devise a ranking algorithm for blog entries that uses the structural characteristic of blogs; the algorithm enables a new blog entry or other entries that have no in-links to be rated according to the past performance of the blogger (2005). There is a fairly recent work closely related to ours performed by Hassan et al (2009) and this work identifies the list of particularly important blogs with recurring interest in a specific topic; their approach is based on lexical similarity and random walks.

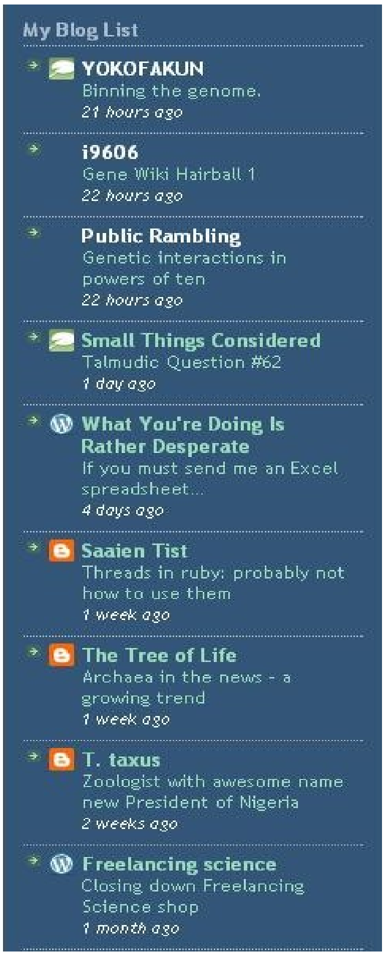
3 Cluster Finding and Ranking Methodology

In this section we explain the concept of "topic clusters" in detail and go into the details of why we deviate from the traditional term of "blog community" in the literature. After this significant discussion we then move on to explain our proposed method for identification and ranking of the "topic clusters" in the blogosphere: two metrics "topic discussion isolation rank" and "topic discussion rank" are used for this purpose.

3.1 Topic Clusters

As explained in section 2 the problem of grouping together blogs has been referred to as the "community detection problem" in the literature. However an aspect ignored by most of these works is the contents of the blogs. Additionally most of the works in this dimension find a blog community by following blog threads' discussions/conversations (Nakajima et al., 2005; Agarwal et al., 2008) which may not always be the case as blogs linking to each other are not necessarily part of communications or threads.

With the advent of micro blogging tools such as Twitter (Honeycutt and Herring, 2009) the role of blogs as a conversational medium has diminished and bloggers link to each other as a socially networked cluster by linking to their most favorite blogs on their home page as is shown in the snapshot of a blog in Figure 1:



My Blog List

YOKOFAKUN
Binning the genome.
21 hours ago

i9606
Gene Wiki Hairball 1
22 hours ago

Public Rambling
Genetic interactions in powers of ten
22 hours ago

Small Things Considered
Talmudic Question #62
1 day ago

What You're Doing Is Rather Desperate
If you must send me an Excel spreadsheet...
4 days ago

Saaien Tist
Threads in ruby: probably not how to use them
1 week ago

The Tree of Life
Archaea in the news - a growing trend
1 week ago

T. taxus
Zoologist with awesome name new President of Nigeria
2 weeks ago

Freelancing science
Closing down Freelancing Science shop
1 month ago

Figure 1: Blog Showing the List of Blogs it Follows

Normally those bloggers link to each other that have similar interests and importantly talk about same topics. Hence the idea of topic cluster is used to extract those clusters from the blogosphere that have strong interest in some specific topics which they mention frequently in their blog posts and additionally they form a linked cluster of blogs. As pointed out by Hassan et al. the "task of providing users with a list of particularly important blogs with a recurring interest in a specific topic is a problem that is very significant in the Information Retrieval domain" (2009). For the purpose of solving this problem we propose the notion of "topic clusters." The task is much different from traditional community detection in the blogosphere as it utilizes both content and link based analysis. The process of finding topic clusters is carried out by calculating a metric "Topic Discussion Isolation Rank" which we explain in detail in section 3.3.

3.2 Rank Assignment to Topic Clusters

As we explained in section 1, due to the unique nature of the blogosphere, traditional link-based methods such as PageRank (Page et al., 1998) may not be appropriate for the ranking task in blogs. This is the main reason that we use the content of blog posts and lexical similarity in blog posts along with links for the rank assignment function that we propose. Furthermore we take a blog as aggregate of all its posts for the retrieval task.

3.3 Topic Discussion Isolation Rank

Topic Discussion Isolation Rank is a metric that is used to find the cluster of blogs for a specific topic. It takes each blog in isolation and analyses the contents of its posts to discover its interest in a queried topic. We consider a blog along three dimensions as Figure 2 shows:

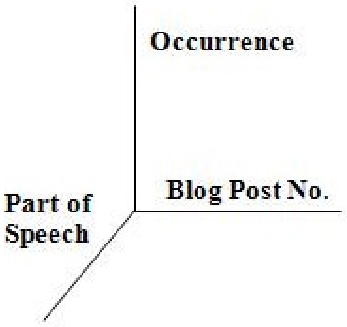


Figure 2: Blog TDIR Dimensions

As mentioned in section 1 of this paper we utilize some natural language processing techniques to ensure better coverage of our cluster-finding and ranking methodology: those techniques are applied along the part of speech dimension shown in Figure 1, for a given topic we analyze blog post contents not only for that particular topic but also for its associated adjectives and adverbs i.e. the topic itself is treated as a noun and its adjectives and adverbs are also used. For example if the topic of interest is "democracy" we will also analyze the blog post contents for adjective "democratic" and adverb "democratically." Furthermore, a weight in descending order is assigned to the noun (denoted as w_n), adjective (denoted as w_{adj}) and adverb (denoted as w_{adv}) of the queried topic where $w_n > w_{adj} > w_{adv}$. This approach guarantees better coverage of the blogosphere and the chances of missing out blogs that have interest in the queried topic are minimal. The blog post number denotes the number of the post in which the word is found and occurrence is a true/false parameter denoting whether or not the word exists in the blog post. Based on these three dimensions we formulated the TDIR metric as follows:

$$1+\frac{(n_{noun} x\ w_n)+(n_{adjective}\ x\ w_{adj})+(n_{adverb} x\ w_{adv})}{Number\ of\ total\ posts}$$

Here $w_{n,}$ w_{adj} and w_{adv} are as explained previously in this section and n_{noun} denotes the number of times nouns are found in all the blog posts, $n_{adjective}$ denotes the number of times adjectives are found in all the blog posts and n_{adverb} denotes the number of times adverbs are found in all the blog posts. This metric is calculated for each blog in isolation and the blogs that have TDIR value of greater than 1 are considered part of the topic cluster.

Additionally we also use various forms of the queried topic in the calculation of TDIR as this also ensures better coverage during the cluster detection process. In the world of the blogosphere, bloggers have all the freedom to use whatever terms they want to use for a particular topic and it is this freedom which adds to the difficulty of the Information Retrieval community. Within the TDIR metric we propose use of alternate terms/spellings/phrases for a given topic – an example being the use of "Obama" by some bloggers and "United States first Black President" or "United States' Black President" by others. Such ambiguity with respect to posts talking about same topic but using different phrases/spellings/terms can be resolved by using a corpus-based approach with listing of alternate phrases and terms for the broad topics.

Moreover the weights used for each of the part of speech “noun”, “adjective” and “adverb” in the TDIR metric can be adjusted differently for different topics with some topics having a stronger indication of discussion of that topic through occurrence of noun and some through occurrence of adjective or adverb. Some examples of these various measures are shown in our experimental evaluations that are explained in section 4.

3.4 Topic Discussion Rank

After the cluster-finding phase we perform the ranking step by means of Topic Discussion Rank. It is in this phase that the socially networked and linked blogs play a role in boosting each other’s ranks. It is reasonable to assign a higher topic rank to a blog that has interest in the specific topic and is also a follower of many blogs with similar topic discussions than one that mentions the topic under consideration but does not link to other similar blogs: Topic Discussion Rank does that by taking into account both link structure and TDIR explained in previous section. This has the advantage of taking into account both factors: the content of the blog posts and the link structure of its neighborhood.

The following piecewise function shows how the metric Topic Discussion Rank is calculated:

$$TDR[b] = \begin{cases} TDIR; \text{ if zero outlinks from blog} \\ TDIR + \frac{Matching\_Outlinks}{Total\_Outlinks} \times \sum_{o:(o,b)} TDIR \times damp; \text{ otherwise} \end{cases}$$

Explanation of notations used:

b - blog

o : (o,b) – outlinks from blog b

The TDR is same as the TDIR in case of the blog having zero outlinks as such a blog exists in isolation and does not have a strong participation within the social network of the blogosphere. In the case of a blog having one or more outlinks to other blogs we add its own TDIR to the factor

$$\left(\frac{Matching\_Outlinks}{Total\_Outlinks} \times \sum_{o:(o,b)} TDIR \times damp\right)$$

.

Here matching links represent blogs that are part of topic cluster for a given topic (i.e. those having TDIR greater than 1 as explained in section 3.3) and each matching link’s TDIR is summed up and multiplied by a factor called damp. Note that summation of TDIR is used in the first iteration only, in the other iterations it is replaced by TDR of the blogs.

Furthermore it is important to note that the process of TDR computation is an iterative one similar to PageRank (Page et al., 1998) computation, however the termination condition is unlike PageRank in that PageRank terminates when rank values are normalized whereas our approach uses the blog depth as a termination condition which is an adjustable parameter. Due to the changed termination condition the role of spam blogs is minimized.

The damping factor *damp* is introduced to minimize biasness as is explained below. Consider the two blogs as shown with the link structure represented by arrows:

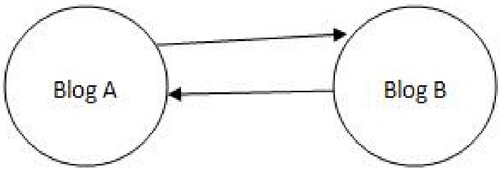


Figure 3: Example for Damping Factor Explanation

In this case let’s assume the TDIR of blog A is 2 and the TDIR of blog B is 1. Using the formulation for TDR without the damping factor we would have 2+(1/1x1)=3 for blog A and 1+(1/1x2)=3 for blog B which is not the true reflection of their topic discussion ranks. However when we use the damping factor the resultant TDR’s are 2+(1/1x1x0.9)=2.9 for blog A and 1+(1/1x2x0.9)=2.8 for blog B and this more correctly represents the topic discussion ranks of both the blogs.

4 Experimental Evaluations

This section presents details of our experiments on real blog data. We use precision and recall to measure the effectiveness of our approach of cluster-finding. The experimental data is released as an XML corpus which can be downloaded from:

http://unhp.com.pk/blogosphereResearch/data.tar.gz.

4.1 Data and Methodology

The data used in the experiments was gathered from 102 blog sites which comprised of 50,471 blog posts. Currently we have restricted the data set to only the blogspot domain (blogger.com service by Google).We used four blog sites as seeds and from them the link structure of the blogs was extracted after which the crawl (Qureshi et al., 2010) was performed using the XML feeds of the blogs to retrieve all the posts in each blog. Each blog had an average of 494 posts.

The topics for which we perform the experiments of finding TDIR and TDR were taken to be *"compute"*, *"democracy"*, *"secularism"*, *"bioinformatics"*, *"haiti"* and *"obama."*

The measures that we use to assess the accuracy of our method are precision and recall which are widely used statistical classification measures for the Information Retrieval domain. The two measures are calculated using equations 4.1 and 4.2:

$$\text{Precision} = \frac{|Ct\ nCa|}{|Ca|} \qquad (4.1)$$

$$\text{Recall} = \frac{|Ct\ nCa|}{|Ct|} \qquad (4.2)$$

Here Ca represents the topic cluster set found using our algorithm i.e. the set of blogs that have interest in the queried topic, in other words it is the set of the blogs that have TDIR greater than 1. Ct represents the true topic cluster set meaning the set of those blogs that not just mention the topic but are really interested in it. The reason for distinguishing between true cluster set Ct and algorithmic cluster set Ca is that our method just searches for the given keyword i.e. topic in all the posts and since natural language is so rich that just mentioning the topic does not represent the fact that the blog is a part of that topic cluster. Hence we use a human annotator/labeler for identification of the true cluster set from the set of the 102 blogs for each of the 6 topics that we used in our experiments.

4.2 Results

We plot the precision and recall graphs for the topics chosen. Figure 4 shows the graph for precision:

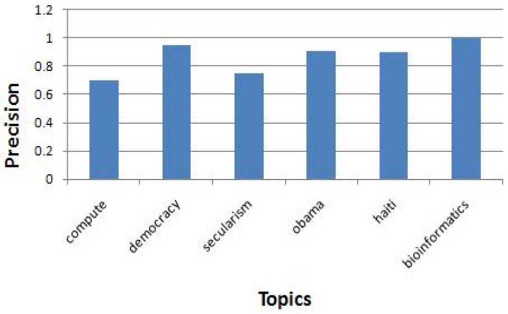


Figure 4: Precision Graph for Chosen Topics

The average precision was found to be 0.87 which reflects the accurate relevance of our method. As can be seen from the graph in figure 4 the precision falls below the 0.8 mark only for the topics compute and secularism – the reason for this is that for these two topics a higher proportion of false positives were discovered. Not all the posts having the word "compute" were actually related to computing as found by human annotator. Same was the case for the word secularism – since our method searches for adjective secular and adverb secularly in case of secularism not being found hence there were some blogs in which secular was used but the blog's focus was not in secularism as an idea. On the other hand precision measures for the topics "democracy", "obama", "haiti" and "bioinformatics" were quite good because these words are likely to be found in the blogs that actually focus on them as a topic hence reducing the chances of false positives.

Figure 5 shows the graph for recall:

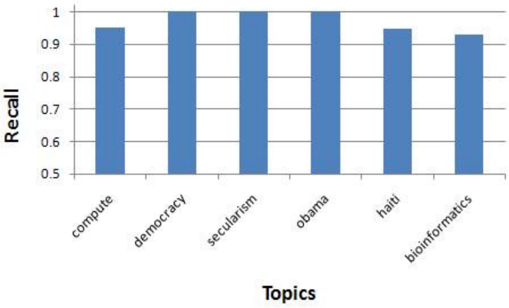


Figure 5: Recall Graph for Chosen Topics

The average recall was found to be 0.971 which reflects the high coverage of our method.

As the graph in figure 5 shows the recall value is mostly close to 1 for the chosen topics. This high coverage is attributed to the part of speech dimension as discussed in section 3.3; this technique rules out the chances of false negatives and hence we obtain a high recall for our method.

4.3 Additional Experiments

In addition to experiments on the six coarse-grained topics mentioned above we performed some additional experiments on two fine-grained topics and also repeated the experiment performed on topic "Obama" with an additional term "Democrats." On formulating the cluster with these two terms the precision increased from 0.907 to 0.95 which clearly shows that incorporation of extra linguistic features into the TDIR formulation ensures better results. Moreover the ranks of some blogs were found to be higher than the ranks obtained previously and this increase in rank was due to the fact that many posts had subject theme "Obama" but they used the term "Democrats" – when we used this alternate term the ranks i.e. TDR more correctly represented the role of the blogs in the cluster.

The two fine grained topics for which we repeated our experiments were: healthcare bill and avatar. Additional terms were also included in the TDIR and TDR computation process which were as follows:

healthcare bill – obamacare
avatar- sky people, jake sully

These alternate terms were chosen as these are the commonly associated terms when these topics are discussed. At this point we provided them as query topics but for future work our plan is to use a machine learning approach for learning these alternate phrases for each topic, and knowledge bases such as Wikipedia may also be used to gather the alternate terms for different topics.

The precision for the topic healthcare bill was found to be 0.857 which had a negligible effect on excluding "obamacare"; however recall suffered more on exclusion of alternate term "obamacare" as it fell from 1 to 0.667. Results for the topic "avatar" however were quite different with a precision of 0.47 and a recall of 1; this was due to the large number of false positives that were retrieved for the term avatar and we found reason for this to be that our approach does not take into consideration case-sensitivity at this point hence it failed to distinguish between the term "avatar" and movie "Avatar". Also in the case of topic "avatar" the alternate phrases did not have any effect and hence there is a need to refine the approach for fine-grained topics such as this one – we present future directions for refinement of our approach in section 5.

5 Conclusions and Future Work

In this paper we proposed the concept of "topic clusters" to solve the blog categorization task for the Information Retrieval domain. The proposed method offers a new dimension in blog community detection and blog ranking by taking into account both link structure and contents of blog posts. Furthermore the natural language processing techniques we use provide a higher coverage thereby leading to a high average recall value of 0.971 in the experiments we performed. At the same time we achieved a good accuracy as was reflected by an average precision of 0.87.

For future work we aim to combine our proposed solution into a framework for auto generation of useful content on a variety of topics such as "blogopedia"; the content can be obtained automatically from the blog posts and in this way manual effort may be saved. We also plan to refine our approach by taking into account the temporal aspects of blog posts such as time interval between blog posts, start post date and time, end post data and time into our formulation for "Topic Discussion Isolation Rank" and "Topic Discussion Rank". Moreover as future directions of this work we plan to incorporate a machine learning framework for the assignment of the weights corresponding to each topic and for the additional phrases to use for each of the topics that we wish to cluster.

References

Agarwal, Nitin, Huan Liu, Lei Tang, and Philip S. Yu, 2008. *Identifying the influential bloggers in a community*. In Proceedings of the international Conference on Web Search and Web Data Mining (Palo Alto, California, USA, February 11 - 12, 2008). WSDM '08. ACM.

Anderson, Paul, 2007. *What is Web 2.0? Ideas, technologies and implications for education.* Technical report, JISC.

Chi, Yun, Shenghuo Zhu, Xiaodan Song, Junichi Tatemura, and Belle L. Tseng, 2007. *Structural and temporal analysis of the blogosphere through community factorization.* In Proceedings of the 13th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining (San Jose, California, USA, August 12 - 15, 2007). KDD '07. ACM.

Fujimura,Ko, Takafumi Inoue, and Masayuki Sugizaki, 2005. *The EigenRumor Algorithm for Ranking Blogs.* In Proceedings of the WWW 2005 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics.

Gruhl, Daniel, R. Guha, David Liben-Nowell, and Andrew Tomkins, 2004. *Information diffusion through blogspace.* In Proceedings of the 13th international Conference on World Wide Web (New York, NY, USA, May 17 - 20, 2004). WWW '04. ACM.

Hassan, Ahmed, Dragomir Radev, Junghoo Cho and Amruta Joshi, 2009. *Content Based Recommendation and Summarization in the Blogosphere*. Third International AAAI Conference on Weblogs and Social Media, AAAI Publications.

Honeycutt, Courtenay, and Susan C. Herring, 2009. *Beyond microblogging: Conversation and collaboration via Twitter.* In Proceedings Hawaii International Conference on System Sciences, IEEE Press

Kumar, Ravi, Jasmine Novak, Prabhakar Raghavan, and Andrew Tomkins, 2003. *On the bursty evolution of blogspace.* In Proceedings of the 12th international Conference on World Wide Web (Budapest, Hungary, May 20 - 24, 2003). WWW '03. ACM.

Leskovec, Jure, Andreas Krause, Carlos Guestrin, Christos Faloutsos, Jeanne Van-Briesen, and Natalie Glance, 2007. *Costeffective outbreak detection in networks*. In The 13th International Conference on Knowledge Discovery and Data Mining (KDD) 2007. ACM.

Li, Hui, Sourav S. Bhowmick, and Aixin Sun, 2009. *Blog cascade affinity: analysis and prediction.* In Proceeding of the 18th ACM Conference on information and Knowledge Management (Hong Kong, China, November 02 - 06, 2009). CIKM '09. ACM.

Mishne, G. and Maarten de Rijke, 2006. *A Study of Blog Search*. In Proceedings of ECIR-2006. LNCS vol 3936. Springer.

Nakajima,Shinsuke, Junichi Tatemura, Yoichiroara Hino, Yoshinori Hara and Katsumi Tanaka, 2005. *Discovering Important Bloggers based on Analyzing Blog Threads.* In Proceedings of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 - 14, 2005). WWW '05. ACM.

O'Reilly, Tim, 2005. *What is Web 2.0: Design Patterns and Business Models for the next generation of software.*

Page, Larry, Sergey Brin, Rajeev Motwani and Terry Winograd, 1999. *The PageRank citation ranking: Bringing order to the Web*, Technical Report, Stanford University.

Qureshi, M. Atif, Arjumand Younus and Francisco Rojas, 2010. *Analyzing Web Crawler as Feed Forward Engine for Efficient Solution to Search Problem in the Minimum Amount of Time through a Distributed Framework.* In Proceedings of 1st International Conference on Information Science and Applications, IEEE Publications.

Sifry, David, 2009 Sifry’s Alerts. http://www.sifry.com/alerts/

Sorapure, Madeleine. 2003. *Screening moments, scrolling lives: Diary writing on the web.* Biography: An Interdisciplinary Quarterly, 26(1), 1-23.

Helping Volunteer Translators, Fostering Language Resources

Masao Utiyama	**Takeshi Abekawa**	**Eiichiro Sumita**	**Kyo Kageura**
MASTAR Project	National Institute	MASTAR Project	Tokyo University
NICT	of Informatics	NICT	
Keihanna Science City	2-1-2, Hitotsubashi	Keihanna Science City	7-3-1 Hongo, Bunkyo-ku
619-0288 Kyoto, Japan	101-8430 Tokyo, Japan	619-0288 Kyoto, Japan	113-0033 Tokyo, Japan
mutiyama@nict.go.jp	abekawa@nii.ac.jp	eiichiro.sumita@nict.go.jp	kyo@p.u-tokyo.ac.jp

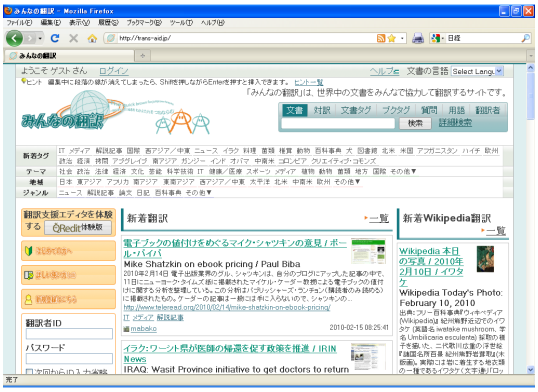
Abstract

This paper introduces a website called *Minna no Hon'yaku* (MNH, "Translation for All"), which hosts online volunteer translators. Its core features are (1) a set of translation aid tools, (2) high quality, comprehensive language resources, and (3) the legal sharing of translations. As of May 2010, there are about 1200 users and 4 groups registered to MNH. The groups using it include such major NGOs as Amnesty International Japan and Democracy Now! Japan.

1 Introduction

This paper introduces a website called *Minna no Hon'yaku* (MNH, "Translation for All", Figure 1), which hosts online volunteer translators (Utiyama et al., 2009).[1] Its core features are (1) a set of translation aid tools, (2) high quality, comprehensive language resources, and (3) the legal sharing of translations.

First, the translation aid tools in MNH consist of the translation aid editor, QRedit, a bilingual concordancer, and a bilingual term extraction tool. These tools help volunteer translators to translate their documents easily as described in Section 3. These tools also produce language resources that are useful for natural language processing as the byproduct of their use as described in Section 4.



みんなの翻訳 - Mozilla Firefox
ファイル(F) 編集(E) 表示(V) 履歴(S) ブックマーク(B) ツール(T) ヘルプ(H)
http://trans-aid.jp/
日経
みんなの翻訳
ようこそ ゲスト さん ログイン
ヘルプ 文書の言語 Select Langu
みんなの翻訳
文書 対訳 文書タグ ブクタグ 質問 用語 翻訳者
検索 詳細検索
新着タグ IT メディア 解説記事 国際 西アジア／中東 ニュース イラク 料理 菌類 椎茸 動物 百科事典 犬 図書館 北米 米国 アフガニスタン ハイチ 欧州 政治 経済 拷問 アブグレイブ 南アジア ガンジー インド オバマ 中南米 コロンビア クリエイティブ・コモンズ
テーマ 社会 政治 法律 経済 文化 芸能 科学技術 IT 健康／医療 スポーツ メディア 植物 動物 菌類 地方 国際 その他▼
地域 日本 東アジア アフリカ 南アジア 東南アジア 西アジア／中東 太平洋 北米 中南米 欧州 その他▼
ジャンル ニュース 解説記事 論文 日記 百科事典 その他▼
翻訳支援エディタを体験する QRedit体験版
はじめての方へ
詳しい使い方はこちら
新規登録はこちら
翻訳者ID
パスワード
□次回からID入力省略
新着翻訳 一覧
新着Wikipedia翻訳 一覧
完了

ヒント 編集中に段落の線が消えてしまったら、Shiftを押しながらEnterを押すと挿入できます。 ヒント一覧

「みんなの翻訳」は、世界中の文書をみんなで協力して翻訳するサイトです。

電子ブックの値付けをめぐるマイク・シャツキンの意見 / ポール・バイバ

Mike Shatzkin on ebook pricing / Paul Biba

2010年2月14日 電子出版業界のグル、シャツキンは、自分のブログにアップした記事の中で、11日にニューヨーク・タイムズ紙に掲載されたマイケル・ケーダー教授による電子ブックの値付けに関する分析を整理している。この分析はパブリッシャーズ・ランチョン（購読者のみ読める）に掲載されたもの。ケーダーの記事は一般には手に入らないので、シャツキンの…

http://www.teleread.org/2010/02/14/mike-shatzkin-on-ebook-pricing/

IT メディア 解説記事

mabako 2010-02-15 08:25:41

イラク：ワーシト県が医師の帰還を促す政策を推進 / IRIN News

IRAQ: Wasit Province initiative to get doctors to return

Wikipedia 本日の写真 / 2010年2月10日 / イワタケ

Wikipedia Today's Photo: February 10, 2010

出典：フリー百科事典『ウィキペディア (Wikipedia)』紀州熊野近辺でのイワタケ (英語名 iwatake mushroom、学名 Umbilicaria esculenta) 採取の様子を描いた、二代歌川広重の浮世絵『諸国名所百景 紀州熊野岩茸取』(木版画)。実際には岩に着生する地衣類の一種であるイワタケ（文字通り「ロッ

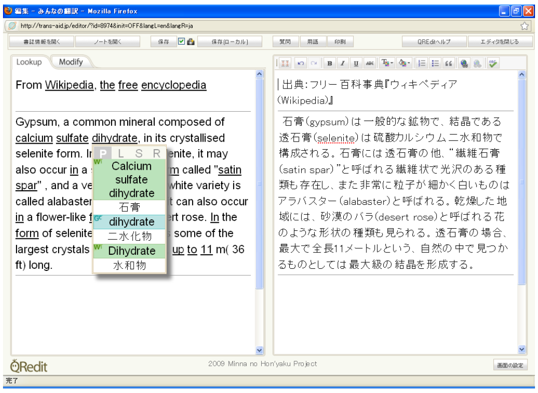
Figure 1: Screenshot of "Minna no Hon'yaku" site (`http://trans-aid.jp`)

Second, MNH provides comprehensive language resources, which are easily looked up in QRedit. MNH, in cooperation with Sanseido, provides "*Grand Concise English Japanese Dictionary*" (Sanseido, 2001) and plans to provide "*Grand Concise Japanese English Dictionary*" (Sanseido, 2002) in fiscal year 2010. These dictionaries have about 360,000 and 320,000 entries, respectively, and are widely accepted as standard and comprehensive dictionaries among translators. MNH also provides seamless access to the web. For example, MNH provides a dictionary that was made from the English Wikipedia. This enable translators to reference Wikipedia articles during the translation process as if they are looking up dictionaries.

Third, MNH uses Creative Commons Licenses (CCLs) to help translators share their translations. CCLs are essential for sharing and opening translations.

[1] Currently, MNH hosts volunteer translators who translate Japanese (English) documents into English (Japanese). The English and Japanese interfaces are available at `http://trans-aid.jp/en` and `http://trans-aid.jp/ja`, respectively.

Beijing, August 2010
Proceedings of the 2nd Workshop on "Collaboratively Constructed Semantic Resources", Coling 2010, pages 63–66,



http://trans-aid.jp/editor/?id=8974&init=OFF&langL=en&langR=ja

From Wikipedia, the free encyclopedia

Gypsum, a common mineral composed of calcium sulfate dihydrate, in its crystallised selenite form. In … enite, it may also occur in a … m called "satin spar" , and a ve… white variety is called alabaster… t can also occur in a flower-like f… rt rose. In the form of selenite… some of the largest crystals … up to 11 m(36 ft) long.

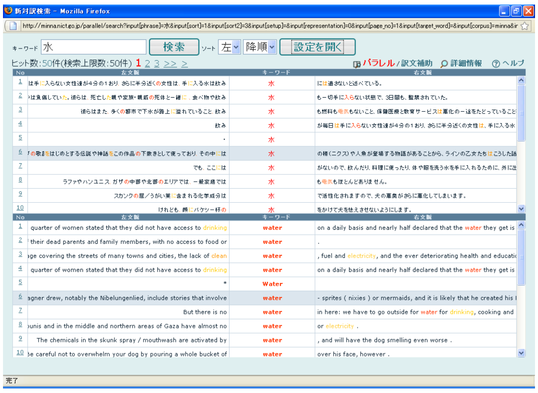
Calcium sulfate dihydrate
石膏
dihydrate
二水化物
Dihydrate
水和物

出典：フリー百科事典『ウィキペディア（Wikipedia）』

石膏（gypsum）は一般的な鉱物で、結晶である透石膏（selenite）は硫酸カルシウム二水和物で構成される。石膏には透石膏の他、“繊維石膏（satin spar）”と呼ばれる繊維状で光沢のある種類も存在し、また非常に粒子が細かく白いものはアラバスター（alabaster）と呼ばれる。乾燥した地域には、砂漠のバラ（desert rose）と呼ばれる花のような形状の種類も見られる。透石膏の場合、最大で全長11メートルという、自然の中で見つかるものとしては最大級の結晶を形成する。

2009 Minna no Hon'yaku Project

Figure 2: Screenshot of QRedit



http://minna.nict.go.jp/parallel/search?input[phrase]=水&input[sort]=1&input[sort2]=3&input[setup]=&input[representation]=0&input[page_no]=1&input[target_word]=&input[corpus]=minna&ir

キーワード 水

ヒット数：50件（検索上限数：50件） 1 2 3 >> >

No	左文脈	キーワード	右文脈
1	は手に入らない女性達が4分の1おり、さらに半分近くの女性は、手に入る水は飲み	水	には適さないと述べている。
2	は負傷していた。彼らは、死亡した親や家族・親戚の死体と一緒に、食べ物や飲み	水	も一切手に入らない状態で、3日間も、監禁されていた。
3	彼らはまた、多くの都市で下水が路上に溢れていること、飲み	水	も燃料も電気もないこと、保健医療と教育サービスは悪化の一途をたどっていること
4	飲み	水	が毎日は手に入らない女性達が4分の1おり、さらに半分近くの女性は、手に入る水
5	・	水	
6	の歌』をはじめとする伝説や神話をこの作品の下敷きとして使っており、その中には	水	の精（ニクス）や人魚が登場する物語があることから、ラインの乙女たちはこうした話
7	でも、ここには	水	がないので、飲んだり、料理に使ったり、体や服を洗う水を手に入れるために、外に出
8	ラファやハンユニス、ガザの中部や北部のエリアでは、一般家庭では	水	も電気もほとんどありません。
9	スカンクの屁／うがい薬に含まれる化学成分は	水	で活性化されますので、犬の悪臭がさらに悪化してしまいます。
10	けれども、顔にバケツ一杯の	水	をかけて犬を怯えさせないようにします。

No	左文脈	キーワード	右文脈
1	quarter of women stated that they did not have access to drinking	water	on a daily basis and nearly half declared that the water they get is
2	their dead parents and family members, with no access to food or	water	.
3	age covering the streets of many towns and cities, the lack of clean	water	, fuel and electricity, and the ever deteriorating health and educati
4	quarter of women stated that they did not have access to drinking	water	on a daily basis and nearly half declared that the water they get is
5	*	Water	
6	agner drew, notably the Nibelungenlied, include stories that involve	water	- sprites (nixies) or mermaids, and it is likely that he created his
7	But there is no	water	in here: we have to go outside for water for drinking, cooking and
8	unis and in the middle and northern areas of Gaza have almost no	water	or electricity .
9	The chemicals in the skunk spray / mouthwash are activated by	water	, and will have the dog smelling even worse .
10	Be careful not to overwhelm your dog by pouring a whole bucket of	water	over his face, however .

Figure 3: Screenshot of bilingual concordancer

2 Related work

There are many translation support tools, such as Google Translator Toolkit, WikiBABEL (Kumaran et al., 2009), BEYtrans (Bey et al., 2008), Caitra (Koehn, 2009) and Idiom WorldServer system,[2] an online multilingual document management system with translation memory functions.

The functions that MNH provides are closer to those provided by Idiom WorldServer, but MNH provides a high-quality bilingual dictionaries and functions for seamless Wikipedia and web searches within the integrated translation aid editor QRedit. It also enables translators to share their translations, which are also used as language resources.

3 Helping Volunteer translators

This section describes a set of translation aid tools installed in MNH.

3.1 QRedit

QRedit is a translation aid system which is designed for volunteer translators working mainly online (Abekawa and Kageura, 2007). When a URL of a source language (SL) text is given to QRedit, it loads the corresponding text into the left panel, as shown in Figure 2. Then, QRedit automatically looks up all words in the SL text. When a user clicks an SL word, its translation candidates are displayed in a pop-up window.

3.2 Bilingual concordancer

The translations published on MNH are used to make a parallel corpus by using a sentence alignment method (Utiyama and Isahara, 2003). MNH also has parallel texts from the Amnesty International Japan, Democracy Now! Japan, and open source software manuals (Ishisaka et al., 2009). These parallel texts are searched by using a simple bilingual concordancer as shown in Figure 3.

3.3 Bilingual term extraction tool

MNH has a bilingual term extraction tool that is composed of a translation estimation tool (Tonoike et al., 2006) and a term extraction tool (Nakagawa and Mori, 2003).

First, we apply the translation estimation tool to extract Japanese term candidates and their English translation candidates. Next, we apply the term extraction tool to extract English term candidates. If these English term candidates are found in the English translation candidates, then, we accept these term candidates as the translations of those Japanese term candidates.

4 Fostering language resources

Being a “one stop” translation aid tool for online translators, MNH incorporates mechanisms which enable users to naturally foster important translation resources, i.e. terminological resources and translation logs.

[2] http://www.idiominc.com/en/

4.1 Terminological resources

As with most translation-aid systems, MNH provides functions that enable users to register their own terminologies. Users can assign the status of availability to the registered terms. They can keep the registered terms for private use, make them available for a specified group of people, or make them publicly available. Several NGO groups are using MNH for their translation activities. For instance, Amnesty International, which uses MNH, maintains a list of term translations in the field of human rights by which translators should abide. Thus groups such as Amnesty upload a pre-compiled list of terms and make them available among volunteers. It is our assumption and aim that these groups make their terminological resources not only available among the group but also publicly available, which will create win-win situation: NGOs and other groups which make their lists of terms available will have more chance of recruiting volunteer translators, while MNH has more chance of attracting further users.

At the time of writing this paper (May 2010), 56,319 terms are registered, of which 45,843 are made publicly available. More than 80 per cent of the registered terms are made public. Currently, MNH does not identify duplicated terms registered by different users, but when the number of registered terms become larger, this and other aspects of quality control of registered terms will become an important issue.

4.2 Translation corpus

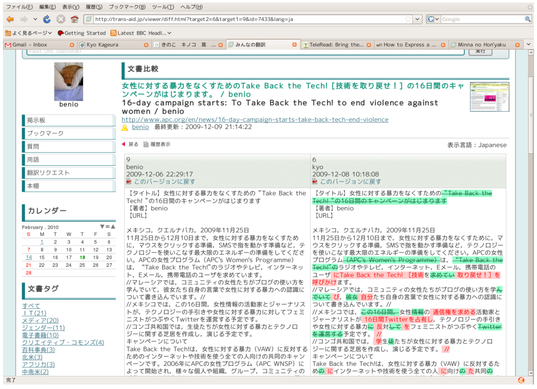
Another important language resources accumulated on MNH is the translation corpus. As mentioned in the introduction, being a hosting site, MNH naturally accumulates source and target documents with a clear copyright status. Of particular importance in MNH, however, is that it can accumulate a corpus that contains draft and final translations made by human together with their source texts (henceforth SDF corpus for succinctness). This type of corpus is important and useful, because it can be used for the training of inexperienced translators (for instance, the MeLLANGE corpus, which contains different versions of translation, is well known for its usefulness in translator training (MeLLANGE, 2009)) and also because it provides a useful information for improving the performance of machine translation and translation-aid systems. While the importance of such corpora has been widely recognized, the construction of such a corpus is not easy because the data are not readily available due to the reluctance on the side of translators of releasing the draft translation data.

The basic mechanisms of accumulating SDF corpus is simple. Translators using MNH save their translations to keep the data when they finish the translation. MNH keeps the log of up to 10 versions of translation for each document. MNH introduced two saving modes, i.e. snapshot mode and normal mode. The translation version saved in the normal mode is overwritten when the next version is saved. Translation versions saved in snapshot mode are retained, up to 10 versions. Translators can thus consciously keep the versions of their translations.

MNH can collect not only draft and final translations made by a single translator, but also those made by different translators. MNH has a function that enables users to give permission for other translators registered with MNH to edit their original translations, thus facilitating the collaborative translations. Such permission can be open-ended, or restricted to a particular group of users.

This function is of particular importance for NGOs, NPOs, university classes and other groups involved in group-based translation. In these groups, it is a common process in translation that a draft translation is first made by inexperienced translators, which is then revised and finalized by experienced translators. If an inexperienced translator gives permission of editing his/her draft translations to experienced translators, the logs of revisions, including the draft and final versions, will be kept on MNH database.

This is particularly important and useful for the self-training of inexperienced translators and thus potentially extremely effective for NGOs and other groups that rely heavily on volunteer



文書比較

女性に対する暴力をなくすためのTake Back the Tech!［技術を取り戻せ！］の16日間のキャンペーンがはじまります。 / benio

16-day campaign starts: To Take Back the Tech! to end violence against women / benio

http://www.apc.org/en/news/16-day-campaign-starts-take-back-tech-end-violence

benio 最終更新：2009-12-09 21:14:22

9 benio 2009-12-06 22:29:17 このバージョンに戻す	6 kyo 2009-12-08 10:18:08 このバージョンに戻す

【タイトル】女性に対する暴力をなくすための ”Take Back the Tech! ”の16日間のキャンペーンがはじまります
【著者】benio
【URL】
 	
メキシコ，クエルナバカ，2009年11月25日
11月25日から12月10日まで，女性に対する暴力をなくすために，マウスをクリックする準備，SMSで指を動かす準備など，テクノロジーを使いこなす最大限のエネルギーの準備をしてください。APCの女性プログラム（APC's Women's Programme）は，“Take Back the Tech!”のラジオやテレビ，インターネット，Eメール，携帯電話のユーザを求めています。
//マレーシアでは，コミュニティの女性たちがブログの使い方を学んでいて，彼女たち自身の言葉で女性に対する暴力への認識について書き込んでいます。//
//メキシコでは，この16日間，女性情報の活動家とジャーナリストが，テクノロジーの手引きや女性に対する暴力に対してフェミニストがつぶやくTwitterを運営する予定です。
//コンゴ共和国では，生徒たちが女性に対する暴力とテクノロジーに関する芝居を作成し，演じる予定です。
キャンペーンについて
Take Back the Tech!は，女性に対する暴力（VAW）に反対するためのインターネットや技術を使う全ての人向けの共同のキャンペーンです。2006年にAPCの女性プログラム（APC WNSP）によって開始され，様々な個人や組織，グループ，コミュニティの
	【タイトル】女性に対する暴力をなくすための~~”Take Back the Tech! ”の16日間のキャンペーンがはじまります~~
	【著者】benio
	【URL】
	
	メキシコ，クエルナバカ，2009年11月25日
	11月25日から12月10日まで，女性に対する暴力をなくすために，マウスをクリックする準備，SMSで指を動かす準備など，テクノロジーを使いこなす最大限のエネルギーの準備をしてください。APCの女性プログラム~~（APC's Women's Programme）~~は，~~“Take Back the Tech!”の~~ラジオやテレビ，インターネット，Eメール，携帯電話のユーザにTake Back the Tech!［技術を~~求めてい~~取り戻せ！］を呼びかけます。
	//マレーシアでは，コミュニティの女性たちがブログの使い方を学~~んでいて~~び，~~彼女~~自分たち自身の言葉で女性に対する暴力への認識について書き込んでいます。//
	//メキシコでは，~~この16日間，~~女性~~情報~~の通信権を求める活動家とジャーナリストが16日間Twitterを占有し，テクノロジーの手引きや女性に対する暴力~~に~~反対~~して~~をフェミニストがつぶやく~~Twitterを運営する~~予定です。//
	//コンゴ共和国では，学生~~徒~~たちが女性に対する暴力とテクノロジーに関する芝居を作成し，演じる予定です。//
	キャンペーンについて
	Take Back the Tech!は，女性に対する暴力（VAW）に反対するため~~の~~にインターネットや技術を使う全ての人に向け~~の~~た共同~~の~~

Figure 4: Comparative view of different translation versions

translators. Many NGOs face chronically the problem of a paucity of good volunteer translators. The retention rate of volunteer translators is low, which increase the burden of a small number of experienced translators, leaving them no time to give advice to inexperienced translators, which further reduce the retention rate of volunteers. To overcome this vicious cycle, mechanisms to enable inexperienced volunteer translators to train themselves in the cycle of actual translation activities is urgently needed and expected to be highly effective. MNH provides a comparative view function of any pairwise translation versions of the same document, as shown in Figure 4. Translators can check which parts are modified very easily through the comparative view screen, which can effectively works as a transfer of translation knowledge from experienced translators to inexperienced translators.

At the time of writing this paper, MNH contains 1850 documents that have more than one translation versions, of which 764 are published. The number of documents translated by a group (more than one translator) is 110, of which 48 are published. Although the number of translations made by more than one translators is relatively small, they are steadily increasing both in number and in ratio.

5 Conclusion

We have developed a website called *Minna no Hon'yaku* (MNH, “Translation for All”), which hosts online volunteer translators. We plan to extend MNH to other language pairs in our future work.

References

Abekawa, Takeshi and Kyo Kageura. 2007. QRedit: An integrated editor system to support online volunteer translators. In *Digital humanities*, pages 3–5.

Bey, Y., K. Kageura, and C. Boitet. 2008. BEYTrans: A Wiki-based environment for helping online volunteer translators. Yuste, E. ed. *Topics in Language Resources for Translation and Localisation*. Amsterdam: John Benjamins. p. 139–154.

Ishisaka, Tatsuya, Masao Utiyama, Eiichiro Sumita, and Kazuhide Yamamoto. 2009. Development of a Japanese-English software manual parallel corpus. In *MT summit*.

Koehn, Philipp. 2009. A web-based interactive computer aided translation tool. In *ACL-IJCNLP Software Demonstrations*.

Kumaran, A, K Saravanan, Naren Datha, B Ashok, and Vikram Dendi. 2009. Wikibabel: A wiki-style platform for creation of parallel data. In *ACL-IJCNLP Software Demonstrations*.

MeLLANGE. 2009. Mellange. `ttp://corpus.leeds.ac.uk/mellange/ltc.`tml.

Nakagawa, Hiroshi and Tatsunori Mori. 2003. Automaic term recognition based on statistics of compound nouns and their components. *Terminology*, 9(2):201–209.

Sanseido. 2001. *Grand Concise English Japanese Dictionary*. Tokyo, Sanseido.

Sanseido. 2002. *Grand Concise Japanese English Dictionary*. Tokyo, Sanseido.

Tonoike, Masatsugu, Mitsuhiro Kida, Toshihiro Takagi, Yasuhiro Sasaki, Takehito Utsuro, and Satoshi Sato. 2006. A comparative study on compositional translation estimation usign a domain/topic-specific corpus collected from the web. In *Proc. of the 2nd International Workshop on Web as Corpus*, pages 11–18.

Utiyama, Masao and Hitoshi Isahara. 2003. Reliable measures for aligning Japanese-English news articles and sentences. In *ACL*, pages 72–79.

Utiyama, Masao, Takeshi Abekawa, Eiichiro Sumita, and Kyo Kageura. 2009. Hosting volunteer translators. In *MT summit*.

Author Index