Data-driven Paraphrasing and Stylistic Harmonization

Gerold Hintz
Research Training Group AIPHES / Language Technology
Department of Computer Science, Technische Universität Darmstadt
hintz@aiphes.tu-darmstadt.de

Abstract

This thesis proposal outlines the use of unsupervised data-driven methods for paraphrasing tasks. We motivate the development of knowledge-free methods at the guiding use case of multi-document summarization, which requires a domain-adaptable system for both the *detection* and *generation* of sentential paraphrases. First, we define a number of guiding research questions that will be addressed in the scope of this thesis. We continue to present ongoing work in unsupervised lexical substitution. An existing supervised approach is first adapted to a new language and dataset. We observe that supervised lexical substitution relies heavily on lexical semantic resources, and present an approach to overcome this dependency. We describe a method for unsupervised relation extraction, which we aim to leverage in lexical substitution as a replacement for knowledge-based resources.

1 Introduction

One of the key research questions in semantic understanding of natural language is bridging the lexical gap; i.e. in absence of lexical overlap between a pair of text segments, judging their semantic content with respect to semantic similarity, entailment, or equivalence. The term *paraphrase* is used to describe semantic equivalence between pairs of units of text, and can be loosely defined as being *interchangeable* (Dras, 1999). Being able to decide if two text units are paraphrases of each other, as well as the reverse direction - generating a paraphrase for a given phrase, are ongoing efforts. Both components are useful in a number of downstream tasks. One guiding use case for the methods developed in the scope of this thesis is their applicability to automatic summarization (Nenkova et al., 2011). In *extractive summarization*, a good summary should select a subset of sentences while avoiding redundancy. This requires detecting semantic equivalences between sentences. *Abstractive summarization* requires a system to further rephrase the summary, to match space constraints, achieve fluency, or unify stylistic differences in multiple source documents. Here, a paraphrasing component can modify the extracted source sentences to meet such external requirements. The primary focus of this work will be the development of novel methods for both *detecting* and *generating* paraphrases of natural language text. In the wider setting of this thesis, we are particularly interested in multi-document summarization (MDS). To scale to the requirements of multi-domain content, our main interest is in knowledge-free and unsupervised methods for these tasks.

The remainder of this paper is structured as follows. In Section 2 we will briefly cover related work in different subareas pertaining to paraphrasing. In Section 3 we will define a number of research questions, which are central to the thesis. Section 4 will then present some ongoing work in lexical substitution and first steps to move towards a knowledge-free unsupervised approach. Finally, Section 5 will give a conclusion and an outlook to future work being addressed in the thesis.

2 Related work

Paraphrase-related research can roughly be categorized into three areas: 1. *Paraphrase identification* - deciding or ranking the degree of how paraphrastic two given elements are; 2. *Paraphrase generation* - given a text element, generate a meaning-preserving reformulation; and 3. *Paraphrase extraction* - given an input corpus, extract meaningful pairs of paraphrastic elements. We will cover each area briefly; an extensive, high-level summary can be found in (Androutsopoulos and Malakasiotis, 2010).

2.1 Paraphrase Identification

The task of paraphrase identification is strongly related to *Semantic Textual Similarity* (STS) and *Recognizing Textual Entailment* (RTE) tasks. STS has most recently been addressed as a shared task at *SemEval-2015* (Agirre et al., 2015), which gives a good overview of current state-of-the art methods. For the specific task of identifying pairs of paraphrases, the use of discriminative word embeddings (Yin and Schütze, 2015) have recently been shown to be effective.

2.2 Paraphrase Generation

Paraphrase generation, being an open generation task, is difficult to evaluate. However, as a preliminary stage to full paraphrasing a number of *lexical substitution* tasks have become popular for evaluating context-sensitive lexical inference since the *SemEval-2007: lexsub* task (McCarthy and Navigli, 2007). A lexical substitution system aims to predict substitutes for a target word instance within a sentence context. This implicitly addresses the problem of resolving the ambiguity of polysemous terms. Over the course of a decade, a large variety of supervised (Biemann, 2013) and unsupervised (Erk and Padó, 2008; Moon and Erk, 2013; Melamud et al., 2015a) approaches have been proposed for this task.

2.3 Paraphrase Extraction

One of the earlier highly successful approaches to paraphrase extraction was shown by Lin and Pantel (2001). The main idea is an extension of the distributional hypothesis from words sharing similar context to similar paths between pairs of words sharing the same substituting words. Thus, a set of similar paths are obtained which can be regarded as prototypical paraphrases. A notable later method to extract a large database of paraphrases makes use of parallel bilingual corpora. The bilingual pivoting method (Bannard and Callison-Burch, 2005) aligns two fragments within a source language based on an overlap in their translation to a "pivoting" language. The paraphrase database, PPDB (Ganitkevitch et al., 2013) was obtained by applying this approach to large corpora.

3 Research Questions

We define a number of research questions (RQ), which have been partially addressed, and shall also provide a guiding theme to be followed in future work.

RQ 1: How can the lexical substitution task be solved without prior linguistic knowledge? Existing approaches to lexical substitution rely frequently on linguistic knowledge. A lexical resource, such as *WordNet* (Fellbaum, 1998), is used to obtain a list of candidate substitutes, and focus is then shifted towards a ranking-only task. State-of-the-art unsupervised systems can still be improved by leveraging lexical resources as candidate selection filters[1]. We argue that this is related mostly to *semantic word relations*. Whereas some semantic relations (synonymy, hypernymy) are well suited for substitution, other relations (antonymy, opposition) are indicators for bad substitutes. Unsupervised distributed methods are susceptible to not recognizing these different word relations, as they still share similar contexts. As part of this research question, we investigate how knowledge-free approaches can be used to overcome this lack of semantic information. We elaborate on this RQ in Section 4.2.

RQ 2: What is the gap between lexical substitution and full paraphrasing? We aim to further examine the remaining gap to a full paraphrasing system that is not restricted to single words. As

[1] We have experimentally confirmed that a fully unsupervised approach (Melamud et al., 2015b) can be improved by restricting substitution candidates to those obtained from *WordNet*

a first step, we extend lexical substitution to multiword expressions (MWE). As most existing research considers only the restricted case of single words, the adaptation of existing features and methods to nominal phrases, and more complex MWEs, will be investigated in detail. Furthermore, the lexical substitution task is conventionally only defined as providing a ranked list of lemmas as target substitutes. In general, directly replacing the target word in the existing context results in a syntactically incorrect sentence. This is is the case for languages with inflection, but also for words with discontinuous expressions, which may require restructuring the sentence. As a next step we plan on leveraging morphological tagging (Schmid and Laws, 2008) to apply syntactic reformulation, by adapting a rule-based transformation framework (Ruppert et al., 2015).

RQ 3: How can a paraphrasing system be employed for stylistic harmonization? In multi-document summarization, source documents frequently originate from different text genres. E.g. a news document employs a different writing style than a blog post or a tweet. Detecting such *stylistic variation* across genres has received some attention (Brooke and Hirst, 2013). Recently, stylistic information has successfully induced for paraphrases within PPDB (Pavlick and Nenkova, 2015). Using a simple log ratio of observation probability of a given phrase across distinct domains, the *style* of the phrase could be mapped in a spectrum for multiple dimensions, such as *formal* / *casual* or *simple* / *complex*. When generating a summary containing such different genres, fluency and coherence of the resulting document have to be considered. To improve summaries, a system could perform the following steps

1. Given an input corpus, *identifying* different styles and given a document *detecting* its style
2. Given an input sentence and its *source style*, paraphrasing it to match a desired *target style*

We can achieve this by considering the difference of distributional expansions across multiple domains. For example, the trigram context “four _ passengers” might frequently be expanded with “aircraft” in a news-domain corpus, whereas a tweet domain more frequently uses “airplane”, with both expansions being distributionally similar. We can thus learn that “aircraft” could be a substitution to adapt towards news-style language and selectively perform such replacements.

RQ 4: Can we exploit structure in monolingual corpora to extract paraphrase pairs? Paraphrase databases, such as PPDB (Ganitkevitch et al., 2013), are constructed from bilingual parallel corpora. Here an assumption is used that equivalent text segments frequently align to the same segment in a different “pivoting” language. The center of this RQ is the goal to extract paraphrase pairs, similar to PPDB, from *monolingual* corpora by exploiting different structure. One such structure can be seen in news corpora. When given a document timestamp, it is possible to exploit the notion of *burstiness* to find out if two documents are related to the same or different events. We aim to adapt techniques aimed at summarization to extract pairs of paraphrases (Christensen, 2015).

4 Ongoing Work and Preliminary Results

4.1 Delexicalized lexical substitution

In a first work we address RQ 1 and perform lexical substitution in a previously unexplored language. With *GermEval-2015* (Miller et al., 2015), the lexical substitution challenge was posed for the first time using German language data. It was shown that an existing supervised approach for English (Szarvas et al., 2013) can be adopted to German (Hintz and Biemann, 2015). Although the wider focus of the thesis will be the use of fully unsupervised methods, in this first step lexical semantic resources are utilized both for obtaining substitution candidates as well as extracting semantic relation features between words. The suitability of various resources, *GermaNet* (Hamp and Feldweg, 1997), *Wiktionary*[2], and further resources crawled from the web, are evaluated with respect to the *GermEval* task. It was found that no other resource matches the results obtained from *GermaNet*, although its coverage is still the primary bottleneck for this system. As lexical substitution data is now available in at least three languages (English, German, and Italian), we also explore language transfer learning for lexical substi-

[2] *Wiktionary*: `https://www.wiktionary.org/`

tution. Experimental results suggest that *delexicalized features* can be extended to not only generalize across lexical items, but can further train a model across languages, suggesting the model to be language independent. For this, we adapt existing features from (Szarvas et al., 2013) and extend the feature space based on more recent approaches. We follow a state-of-the-art unsupervised model (Melamud et al., 2015b) to further define features in a syntactic word embedding space. As a preliminary result, feature ablation tests show that the strongest features for lexical substitution are semantic relations from multiple aggregated lexical resources. This insight motivates the next step towards a knowledge-free system.

4.2 Unsupervised semantic relation extraction

Semantic relations have been identified a strong features for lexical substitution (Sinha and Mihalcea, 2009); selecting candidates based on aggregated information of multiple resources usually results in good performance. Consequently, when obtaining substitution candidates from different sources, such as a distributional thesaurus (DT), a key challenge lies in overcoming a high amount of *related* but not substitutable words. Prime examples are *antonyms*, which are usually distributionally similar but no valid lexical substitutions (replacing "hot" with "cold" alters the meaning of a sentence). Figure 1 illustrates this challenge at the example of an instance obtained from the SemEval-2007 data. Here, candidates from a DT are compared against candidates obtained from *WordNet*. Both resources yield related words (e.g. "task", "wage", "computer science" are all related to the target "job") - however, for lexical resources we can leverage semantic relations as a much more fine-grained selection filter beyond relatedness (in the example, entries such as "computer science" can be excluded by discarding the *topic* relation). On the other hand, obtaining candidates only from a lexical resource necessarily limits the system to its coverage. Whereas *WordNet* is a high-quality resource with good coverage, alternatives for other languages may be of inferior quality or are lacking altogether. To quantify this, we have evaluated the overlap of semantic relations present in *GermaNet* (Hamp and Feldweg, 1997) with the gold substitutes in the *GermEval-2015* lexsub task. Figure 2 illustrates this overlap and further shows the stages at which the resource fails. Whereas all target words are contained in GermaNet, only 85% of the substitutes are contained as lexical items. When considering only *synonyms*, *hyponyms* and *hypernyms* as relations, only 20% of all gold substitutes can be retrieved. This number constitutes an upper bound for the recall of a lexical substitution system. If, instead, candidates are obtained based on distributional similarity, we can obtain a much higher upper bound on recall of substitution candidates. Figure 3 shows the recall of the top-k similar words, based on a distributional thesaurus computed from a 70M sentence newspaper corpus (Biemann et al., 2007). Even when considering only the top-50 most similar words, a recall of 29% can be achieved, whereas this value plateaus at about 45% - improving over the lexical resource baseline more than twofold. In summary, we make two observations:

1. Similarity-based approaches, such as distributional similarity, have better coverage for substitution candidates, at the cost of higher noise

His **job** was unpaid, but he was working just to keep fit.
work (4)
employment (2)
post (1)

DT entries	WordNet entries	(label)
job#NN	business	synset
task#NN	occupation	synset
employment#NN	task	co-hyponym
position#NN	chore	co-hyponym
worker#NN	activity	hypernym
post#NN	**work**	hyponym
employee#NN	**employment**	hyponym
acre#NN	**post**	hyponym
foot#NN	obligation	hypernym
wage#NN	computer science	topic
work#NN	computing	topic

Figure 1: Comparison of substitution candidates obtained from a DT (most similar words) and a lexical resource (WordNet), for a given example sentence from SemEval-2007. Bold items denote overlap with gold substitutes.

$t \in L$ 100%	$s \in L$	85%	$\exists R.R\,(t,s)$	20%	syn	5%
					hypo	8%
					hyper	16%
			$\neg\exists R.R\,(t,s)$	64%		
	$s \notin L$	15%				

Figure 2: Presence of lexical substitution gold pairs in a lexical resource L. $t \in L$ denotes that a target is present in L. $\exists R.R\,(t,s)$ denotes the fraction within those pairs for which a semantic relation existed. We used GermaNet as a lexical resource L and compare to gold substitutes from *GermEval-2015*
.

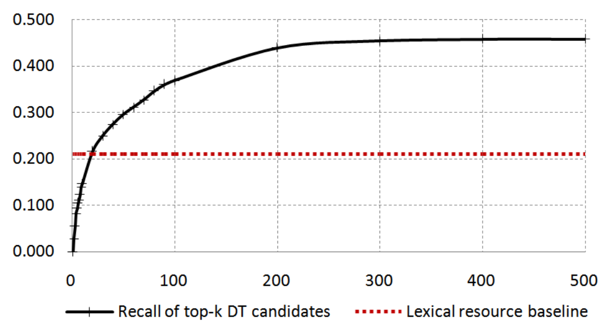

Figure 3: Recall of lexical substitution candidates as top-k similar DT entries, compared to lexical resource baseline (using synonyms, hyponyms and hypernyms)

2. The *semantic relation* between target and substitute is a strong indicator for substitutability

These observations motivate a similarity-based selection of substitution candidates, which does not rely on knowledge-based resources. We argue that the second key component to lexical substitution is an unsupervised extraction of semantic relations. For this we follow (Herger, 2014), who leverages the *extended distributional hypothesis*, stating that "if two *paths* tend to occur in similar context, the meanings of the *paths* tend to be similar" (Lin and Pantel, 2001). The original motivation for this is obtaining inference rules, or equivalences between paths. For example, it can be discovered that "*X is author of Y*" $\approx$ "*X wrote Y*". In reverse however, we can also discover pairs of words (X, Y), which tend to be connected with the same paths. We can thus compute a DT on pairs of words rather than single words, using their path as context features. Our al-

After the bubble burst, prices plunged and demand vanished.

$\downarrow$ extract pairs

(bubble, burst)

(price, demand)

$\downarrow$ compute path features

(price, demand) $\rightarrow$ `X plunged and Y`

(price, demand) $\rightarrow$ `X-cc_plunge_Y`

(price, demand) $\rightarrow$ `X-cc_#VB_Y`

Figure 4: Extraction of pairs and path context features. Context features shown here are token substring, lemmatized syntactic dependency edges, and POS-normalized dependency edges

gorithm for semantic relation extraction can thus be described as follows:

1. Extract pairs of words from background corpus, based on distributional similarity
2. Compute context features for each pair based on their connecting path
3. Compute the similarity between pairs of pairs, based on shared context features
4. Cluster pairs of words based on their similarity

For step 1, we experimented with different filtering strategies, based on word frequency, sentence length, token distance, and POS tags. As context features we aggregate multiple variants to generalize a path: we replace the occurrence of the pair (X, Y) with the literal strings "X" and "Y" and then extract the token substring, as well paths defined by syntactic dependency edges. Steps 1 and 2 are visualized in Figure 4. For steps 3 and 4, we adapt the DT computation from (Biemann and Riedl, 2013) and obtain similarity scores based on the overlap of the most salient context features, i.e. generalized paths. At this stage, we obtain a distributional similarity between pairs of words, e.g. the pair *(famine, epidemic)* is distributionally similar to (*problem*, *crisis*). This resembles the notion of word analogies, which can be obtained in embedding spaces (Levy et al., 2014), however our model results in discrete notions of relations as opposed to non-interpretable vectors. For this, we cluster word pairs by applying Chinese Whispers (Biemann, 2006). Table 1 shows the final output of the semantic relation clustering exemplified for four resulting clusters. Although the data is not perfectly consistent, clusters tend to

Cluster 1	Cluster 2
painter::designer	sailboat::boat
welder::electrician	trawler::boat
architect::engineer	ship::boat
sailor::pilot	helicopter::boat
poet::artist	helicopter::vessel
pull::push	coat::dress
consultant::specialist	plane::vessel
distributor::producer	soldier::policeman
decorator::gardener	driver::passenger
Cluster 3	**Cluster 4**
rain::drought	glaucoma::blindness
heat::cold	exposure::illness
legroom::mobility	famine::malnutrition
concern::anger	traffic::pollution
exercise::eating	humans::stress
vengeance::forgiveness	obesity::illness
competitiveness::efficiency	overwork::depression
respect::contempt	hurricane::flooding
hipness::authenticity	inflammation::pain
supervision::management	drought::crisis

Table 1: Exemplary output of semantic relation clustering (cluster subsets)

represent a similar relation between each respective pairs of words. These relations often correspond to those found in lexical resources, such as hyponymy or antonymy. However, the relations are frequently fragmented into smaller, domain-specific clusters. In the above example, Cluster 1 and Cluster 2 both correspond to a relation resembling *hypernymy* - however, in Cluster 1 this relation is mostly clustered for professions (e.g. a "welder" is-a-kind of "electrician"), whereas Cluster 2 corresponds roughly to vehicles or items (a "sailboat" is-a-kind-of "boat"). Cluster 3 can be reasonably considered as containing antonymous terms ("rain" is-opposite-of "drought"). In some cases, clusters contain relations of words not generally found in semantic resources. Cluster 4 contains word pairs having a causation relation (e.g. "glaucoma" *causes* "blindness"); it is further interesting to observe that items in this cluster contain exclusively negative outcomes ("illness", "stress", "flooding", etc.). Previous work has conventionally evaluated semantic relation extraction intrinsically with respect to a lexical resource as a gold standard (Panchenko and Morozova, 2012). However, we are interested in utilizing semantic relations for paraphrasing tasks and will therefore follow up with an extrinsic evaluation in a lexical substitution system. Our goal is to leverage unsupervised clusters, e.g. as feature input, to overcome the need for lexical semantic resources.

5 Conclusion and Outlook

In this paper we have outlined the guiding theme of a thesis exploring data-driven methods for paraphrasing and defined a set of research questions to sketch a path for future work. We addressed the first step of lexical substitution. We showed that a supervised, delexicalized framework (Szarvas et al., 2013) can be successfully applied to a previously unexplored language. We make a number of observations on multiple language lexsub tasks: Obtaining substitution candidates from lexical resources achieves best system performance, despite incurring a very low upper bound on substitution recall. Obtaining candidates in an unsupervised manner by considering distributionally similar words increases this upper bound more than twofold, at the cost of more noise. We further observe that the strongest features in this setting are semantic relations between target and substitute, obtained from the aggregated lexical resources. Hence, we conclude that obtaining semantic relations in an unsupervised way is a key step towards knowledge-free lexical substitution. We continue to present an unsupervised method for obtaining clusters of semantic relations, and show preliminary results. As a next step we aim at integrating such relation clusters into a lexical substitution system. We also plan on extending lexical substitution towards a full paraphrasing system, by moving from single-word replacements to longer multiword expressions, as well as applying syntactic transformations as a post-processing step to the substitution output. In related branches of this thesis we will also explore methods for extracting paraphrases from structured corpora, and ultimately apply a two-way paraphrasing system to a multi-document summarization system, supporting both selection of non-redundant sentences as well as sentence rephrasing to perform harmonization of language style.

Acknowledgments

This work has been supported by the German Research Foundation as part of the Research Training Group "Adaptive Preparation of Information from Heterogeneous Sources" (AIPHES) under grant No. GRK 1994/1.

References

Eneko Agirre, Carmen Banea, et al. 2015. Semeval-2015 task 2: Semantic textual similarity, English, Spanish and pilot on interpretability. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, Denver, CO, USA.

Ion Androutsopoulos and Prodromos Malakasiotis. 2010. A survey of paraphrasing and textual entailment methods. *Journal of Artificial Intelligence Research*, 38:135–187.

Colin Bannard and Chris Callison-Burch. 2005. Paraphrasing with bilingual parallel corpora. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 597–604, Ann Arbor, MI, USA.

Chris Biemann and Martin Riedl. 2013. Text: Now in 2D! a framework for lexical expansion with contextual similarity. *Journal of Language Modelling*, 1(1):55–95.

Chris Biemann, Gerhard Heyer, Uwe Quasthoff, and Matthias Richter. 2007. The leipzig corpora collection: monolingual corpora of standard size. In *Proceedings of Corpus Linguistics*, Birmingham, UK.

Chris Biemann. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the first workshop on graph based methods for natural language processing*, pages 73–80, New York, NY, USA.

Chris Biemann. 2013. Creating a system for lexical substitutions from scratch using crowdsourcing. *Language Resources and Evaluation*, 47(1):97–122.

Julian Brooke and Graeme Hirst. 2013. A multi-dimensional bayesian approach to lexical style. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational*, pages 673–679.

Janara Maria Christensen. 2015. *Towards Large Scale Summarization*. Ph.D. thesis, University of Washington.

Mark Dras. 1999. *Tree Adjoining Grammar and the Reluctant Paraphrasing of Text.* Ph.D. thesis, Macquarie University.

Katrin Erk and Sebastian Padó. 2008. A structured vector space model for word meaning in context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 897–906, Waikiki, HI, USA.

Christiane Fellbaum. 1998. *WordNet*. Wiley Online Library.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *HLT-NAACL*, pages 758–764, Atlanta, GA, USA.

Birgit Hamp and Helmut Feldweg. 1997. GermaNet - a lexical-semantic net for German. In *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, pages 9–15.

Priska Herger. 2014. Learning semantic relations with distributional similarity. Master's thesis, TU Berlin.

Gerold Hintz and Chris Biemann. 2015. Delexicalized supervised German lexical substitution. In *Proceedings of GermEval 2015: LexSub*, pages 11–16, Essen, Germany.

Omer Levy, Yoav Goldberg, and Israel Ramat-Gan. 2014. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 171–180, Baltimore, MD, USA.

Dekang Lin and Patrick Pantel. 2001. DIRT - discovery of inference rules from text. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 323–328, San Francisco, CA, USA.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 48–53, Prague, Czech Rep.

Oren Melamud, Ido Dagan, and Jacob Goldberger. 2015a. Modeling word meaning in context with substitute vectors. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 472–482, Denver, CO, USA.

Oren Melamud, Omer Levy, Ido Dagan, and Israel Ramat-Gan. 2015b. A simple word embedding model for lexical substitution. *VSM Workshop*. Denver, CO, USA.

Tristan Miller, Darina Benikova, and Sallam Abualhaija. 2015. GermEval 2015: LexSub – A shared task for German-language lexical substitution. In *Proceedings of GermEval 2015: LexSub*, pages 1–9, Essen, Germany.

Taesun Moon and Katrin Erk. 2013. An inference-based model of word meaning in context as a paraphrase distribution. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 4(3):42.

Ani Nenkova, Sameer Maskey, and Yang Liu. 2011. Automatic summarization. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts of ACL 2011*, page 3, Portland, OR, USA.

Alexander Panchenko and Olga Morozova. 2012. A study of hybrid similarity measures for semantic relation extraction. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 10–18, Jeju, South Korea.

Ellie Pavlick and Ani Nenkova. 2015. Inducing lexical style properties for paraphrase and genre differentiation. In *Proceedings of the 2015 Annual Conference of the North American Chapter of the Association for Computational Linguistics, NAACL*, pages 218–224, Denver, CO, USA.

Eugen Ruppert, Jonas Klesy, Martin Riedl, and Chris Biemann. 2015. Rule-based dependency parse collapsing and propagation for German and English. In *Proceedings of the German Society for Computational Linguistics & Language Technology 2015*, pages 58–66, Essen, Germany.

Helmut Schmid and Florian Laws. 2008. Estimation of conditional probabilities with decision trees and an application to fine-grained pos tagging. In *Proceedings of the 22nd International Conference on Computational Linguistics*, volume 1, pages 777–784, Manchester, UK.

Ravi Sinha and Rada Mihalcea. 2009. Combining lexical resources for contextual synonym expansion. In *Proceedings of the Conference in Recent Advances in Natural Language Processing*, pages 404–410, Borovets, Bulgaria.

György Szarvas, Chris Biemann, Iryna Gurevych, et al. 2013. Supervised all-words lexical substitution using delexicalized features. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1131–1141, Atlanta, GA, USA.

Wenpeng Yin and Hinrich Schütze. 2015. Discriminative phrase embedding for paraphrase identification. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1368–1373, Denver, CO, USA.