Extracting Professional Preferences of Users from Natural Language Essays

Cigdem TOPRAK [1], Christof MÜLLER and Iryna GUREVYCH
Ubiquitous Knowledge Processing (UKP) Lab, Computer Science Department
Technical University of Darmstadt, Germany
www.ukp.tu-darmstadt.de

Abstract. This paper presents an unsupervised sentiment analysis approach for extracting professional preferences of users from natural language essays in a career recommendation scenario. Our system first extracts terms facilitating career recommendation such as objects, activities, hobbies, and places from the essays. Then, it applies a lexicon-based sentiment analysis approach to assign polarities representing user preferences.

Keywords. unsupervised sentiment analysis, information extraction

Introduction

Subjectivity and sentiment analysis are computational linguistics tasks focusing on the automatic analysis of subjective content in text. Subjectivity analysis aims at automatically distinguishing subjective content (opinions) from objective content (factual information). Sentiment analysis, on the other hand, involves additional subtasks such as: *(i)* determining the emotional orientation (polarity) of the subjective content, i.e., determining whether the analysed content conveys a positive, negative or neutral attitude towards its target, and *(ii)* determining the targets of the opinions.

Recently, subjectivity and sentiment analysis gained an increasing importance as they support information retrieval (IR) and information extraction (IE) applications especially in the domains containing a vast amount of subjective content such as humanities. For instance, subjectivity and sentiment analysis can support IR in two possible ways: *(i)* in query preprocessing where the user is allowed to enter complex natural language queries, such as `negative criticism about the works of A`, subjectivity and sentiment analysis components can help the system to classify this query as opinionated with negative polarity towards the target `works of A`; *(ii)* in opinion-oriented information retrieval, knowing that the query is opinionated and negative, the IR system can retrieve the documents containing opinionated snippets with negative polarity towards the target in the query.

In this paper, we present an unsupervised sentiment analysis component, and its intrinsic evaluation for facilitating query preprocessing in a semantic IR system for career

[1]Corresponding Author: Cigdem Toprak, UKP Lab, Technical University of Darmstadt, Hochschulstr. 10, 64289 Darmstadt, Germany; E-mail: c_toprak@tk.informatik.tu-darmstadt.de

recommendation [1]. The overall system allows users to describe their interests in short essays, called *professional profiles*, treated as queries in IR. Example 1 presents a sample profile.

> **Example 1.** I would like to work with animals, to treat and look after them, but I cannot stand the sight of blood and take too much pity on the sick animals. On the other hand, I like to work with computer, can program in C, Python and VB and so I could consider working in software development. I cannot imagine working in a kindergarden, as a social worker or as a teacher, as I am not very good at asserting myself.

Professional profiles contain words or phrases describing objects `(computer)`, places `(kindergarden)`, activities `(program)`, and profession names `(teacher)` which help the IR system to pinpoint suitable professions for the user based on textual descriptions of professions contained in a database. However, as the example illustrates, professional profiles contain both preferred `(computer, program)` and dispreferred `(kindergarden, social worker)` items. Our sentiment analysis component aims at extracting *professional preferences* which are modelled as *(target, polarity)* tuples where *polarity* reflects user's preference regarding the *target*. This way terms with negative polarities can be excluded from the actual IR query.

The remainder of this paper is structured as follows: We describe the corpus of professional profiles and the evaluation gold standards in Section 2. Sections 3 and 4 present our approaches to target extraction and sentiment analysis as well as the result analysis of both tasks. Finally, we draw some conclusions in Section 5.

1. Annotation Study

The professional profiles corpus contains 45 profiles consisting of 311 sentences and 4668 words in total. We collected 30 profiles from university students and the rest from high school students. We used 30 profiles as the development set and 15 profiles for testing purposes. We manually annotated *professional preferences* using the annotation scheme presented in the next subsection.

1.1. Annotating Professional Preferences

The manual annotation of the *professional preferences* requires annotators to first mark *targets* of the *professional preferences*, i.e., mark spans of words without any part-of-speech restrictions, and then, assign the *category* and *polarity* values to the marked *target*.

While defining the *professional preference* notion, we considered the fact that the extracted *professional preferences* were going to be used by an IR system, not a human. For instance, consider the sentence in Example 2 where the user prefers a challenging job.

> **Example 2.** In any case my future job should challenge me and should not be boring, and I want to be able to realize my own ideas.

While humans can make sense of this sentence in the career recommendation process, it is not informative enough for an IR system as it describes meta-characteristics of a desired profession rather than concrete objects, places, or activities involved in the profession. Therefore, we define the *targets* of *professional preferences* as *objects*, *activities*, *places*, and *profession names* or *areas*, i.e, as concrete clues which define a job or differentiate one job from another. The *category* attribute captures the types mentioned in our definition with the possible values of *object, activity, place, profession*, or *other*. *Other* value is used to mark the targets not fitting within any given category.

The *polarity* attribute represents the user's preference regarding the marked *target* with the possible values of *positive, negative* or *neutral*. Table 1 shows example annotations of the *professional preferences* for the profile presented in Example 1.

Professional preference	Category	Polarity
sick animals, blood	object	negative
animals, computer	object	positive
C, Python, VB	object	positive
program	activity	positive
software development	profession	positive
kindergarden	place	negative
social worker, teacher	profession	negative

Table 1. Example professional preference annotations

The polarity attribute is assigned the *positive* or *negative* value if there is an explicit mention of a preference or dispreference regarding the target. For instance, `like to work with`, and `cannot imagine working` in Example 1 illustrate explicit mentions of preference and dispreference respectively. The *neutral* value is used for the cases where the user mentions a target without indicating an explicit preference. For instance, the polarity for the target `photo laboratory` in the sentence, `I worked in a photo laboratory once`, would be annotated as *neutral* as it does not contain an explicit cue for a specific preference.

1.2. Inter-Annotator Agreement Study

Two linguistics students annotated 28 profiles from the development corpus according to the described scheme. They were given two profiles for training. The annotation process requires marking word spans. Therefore, the annotations exhibit variations in word length. We considered two annotations to be matches if one is a subset of the other in terms of word spans and they intend to mean the same *target*. For instance, in Example 3 the spans `testing improvements` and `testing` are counted as matches since they refer to the same activity.

Example 3. I enjoy transferring knowledge to machines, and then [[testing]$_{AnnA}$ improvements].[2] $_{AnnB}$

The number of the *professional preferences* identified by two annotators differs. For instance, consider the sentence in Example 4 where annotator A marked 5, and annotator B marked 3 targets.

[2] Es macht mir Spaß einer Machine Wissen zu vermitteln, und dann die Fortschritte zu testen.

Example 4. Maybe I could work [in the [industry]$_{AnnA}$ in a [big corporation]$_{AnnA}$ in the [executive board]$_{AnnA}$]$_{AnnB}$, for example at [Daimler]$_{AnnA,AnnB}$ (I like [cars]$_{AnnA,AnnB}$).[3]

For *expression level* agreement calculation we used the directional metric *agr* as proposed in [2]. Let A and B be two sets of annotations marked by two annotators A and B, agreement of annotator B to annotator A is measured as:

$$agr(A||B) = \frac{|A\, matching\, B|}{|A|} \tag{1}$$

The directional metric *agr* measures what proportion of A was also annotated by the annotator B. In other words, $agr(A\|B)$ corresponds to recall if B is being evaluated and A is the gold standard, and to precision if B is the gold standard and A is being evaluated. We obtained an $agr(A\|B)$ of 0.76 and $agr(B\|A)$ of 0.83 where $|A\, matching\, B|$ was 256, i.e., we considered 256 annotations from both sets as refering to the same targets according to the previously mentioned matching criteria. For the 256 matching *professional preferences* we reach a kappa of 0.87 and 0.68 for the *category* and the *polarity* attributes respectively. Based on the sufficient agreement in marking *targets*, *category* and *polarity*, only one annotator labelled the test corpus.

2. Target Extraction

The majority of the *targets* of *professional preferences* consist of nouns, noun phrases and verbs. However, extracting all nouns and verbs as *targets*, despite prunning efforts via a stop list, results in an overgeneration of *professional preferences* as reported by [3]. Therefore, we apply a two-stage approach for extracting the *targets* of the *professional preferences*. We first mark the words belonging to a *target* using an automatically generated lexicon, hereafter called a *target constituent lexicon*. Then, we apply a set of manually defined extraction patterns on the marked words *(target constituents)* to finalize the extraction.

Marking target constituents: *Target constituents* are words which make up the *target*. For instance, the words `English` and `teacher` are the *target constituents* of the *target* `English teacher`. We populate a *target constituent lexicon* from GermaNet [4] using a seed term list of 57 words, whereby 41 words were collected from the tagset of the BERUFEnet portal[4] and 16 words were artificial concepts[5] from GermaNet. The tagset contains three categories of keywords describing *objects* (e.g. `media, foreign language, food`), *places* (e.g. `zoo, garden, manufacturing plant`), and *activities* (e.g. `plant, paint, build`). We retrieved the synsets for

[3] Vielleicht könnte ich in der Wirtschaft bei einem grossen Unternehmen im Vorstand arbeiten, zum Beispiel bei Daimler (ich mag nämlich Autos)

[4] http://interesse-beruf.de provided by the German Federal Labor Office

[5] Artificial concepts represent unlexicalized concepts in the language. For example, *selbständiger_Mensch* and *angestellter_Mensch* are artificial concepts which are co-hyponyms of the *Mensch* concept, however they do not represent real lexical items. Our list of artificial concepts include *Schultyplehrer, hierarchischer_Lehrer, funktionaler_Lehrer, berufstätiger_Mensch, selbstständiger_Mensch, angestellter_Mensch, hausangestellter_Mensch, Strassenberufler, Heilberufler, Sozialberufler, Medienberufler, Lebensmittelverarbeiter, verbeamteter_Mensch, professioneller_Mensch, ausgebildeter_Mensch, abgeordneter_Mensch*

each seed term from GermaNet, and then, recursively queried each sense for its hyponyms in GermaNet. Thereby, we preserved the category of the seed term and assigned it as the category of the terms generated based on the seed term. We apply this approach until we observe no change in the size of the resulting lexicon. Furthermore, we used the 16 seed terms from artificial concepts for populating terms of the *profession* category. As a result, we obtained a lexicon of approximately 9000 terms with category assignments.[6]. We utilized the *target constituent lexicon* to mark the *target constituents*. Additionally, we marked named entities, words tagged with *NE* tag by the POS tagger, as the *target constituents*.

Extraction step: We perform POS tagging and chunking using TreeTagger [5]. Then, we apply the following extraction rules based on POS and chunk tags:

1. **Base noun phrase pattern** extracts base noun phrases containing the *target constituents*. Phrases conform to the POS pattern $(ADJ\ NN^* \| NN^* \| NE^*)$ where NN and NE are the noun and named entity *target constituents*. Example targets extracted according to this pattern include `sick animals`, `electrical equipment`.
2. **Infinite verb pattern** analyses consecutive NC (noun phrase) and VC (verb phrase) chunks where the NC chunk contains a *target constituent* and the VC chunk contains a construction (*to*[7] *infinite verb*). We first apply the *base noun phrase pattern* to the NC chunk and then add the infinite verb from the VC chunk to the base noun phrase. Example *target* extractions based on this pattern include `transfer knowledge to machines`[8], `see other countries`[9].
3. **Coordination pattern** analyses coordinated NC chunks which are the constituents of the *and* and *or* coordinations. According to this pattern, we extract the base noun phrase in a coordinated chunk as a *target*, if the other coordinated chunk already contains a *target*. Example *targets* extracted based on this pattern include `IT Sector` from the sentence `I'm interested in languages and IT Sector` where `languages` was already extracted as a target.
4. **Prepositional phrase pattern** extracts *targets* from PC chunks (prepositional phrases) headed by *with, in*, and *at*[10]. We apply the *base noun phrase pattern* to such PC chunks even if they do not include any *target constituents*. An example *target* extraction based on this pattern includes `software development` from the sentence `I see my future in software development.`

We evaluate our system against a human annotator who marked 336 targets in the development and 96 targets in the test corpus. We used the same matching strategy applied in the expression-level inter-annotator agreement study. We obtain a precision of 0.41, and a recall of 0.66 on the development, and a precision of 0.40, and a recall of 0.55 on the test corpus. Except for the *prepositional phrase pattern* which is lexicon independent, all extraction patterns rely on lexicon look-ups. Therefore, both the coverage and the quality of the *target constituent lexicon* play a crucial role in our extraction approach.

[6]The feature lexicon contains duplicates due to the fact that different seed terms from different categories occasionally populated the same terms. We kept both terms with different category assignments

[7]German: zu

[8]German: Machine Wissen vermitteln

[9]German: andere Länder sehen

[10]with: (German) mit, in: (German) in, at: (German) bei

However, the quality of an automatically generated lexicon is suboptimal as it contains too many noisy terms.

We observe that most of the unextracted *target constituents* were due to the poor coverage of the lexicon, whereas spurious extractions resulted from the overgeneration. For instance, the *targets* `wine` and `nutritional science` in the sentence `By the way, my interests include`[11] `wine and nutritional science` were not extracted due to insufficient coverage. On the other hand, we extracted `interests`[12] as a target due to the noise in lexicon.

In the profiles, users describe their professional interests, dislikes and expectations without any restrictions. We observe that the *targets* of the *professional preferences* are not restricted to those included in a lexicon created from a small set of seed terms.

3. Sentiment Analysis

In the sentiment analysis stage, we assign polarities to the extracted targets, in other words, we focus on the preference part. We utilize the *opinion lexicon* from [3] which contains unigrams with manually assigned polarities. The *opinion lexicon* is generated from GermaNet using the hyponyms of the concepts *feeling (Gefühl)*; *to feel, to care (empfinden)*, and the artificial concepts *evaluation specific (Bewertungsspezifisch)*, *feeling specific (Gefühlspezifisch)*. We utilize the lexicon to mark the *sentiment cues*.

We perform full parsing with BitPar[13]. BitPar delivers parse trees with nodes representing syntactic categories (e.g. S: sentence, NP: noun phrase, VP: verb phrase, PP: prepositional phrase), and edges representing functional units (e.g. SB: subject, HD: head, PD: predicate) and modification relations (e.g. MO: modifier, NG: negation).

We analyse each clause as a separate unit based on the assumption that a clause represents a unit of thought. We assign polarities on the clause level. A clause can be identified as an S node headed by a finite verb in a parse, or a VP node as a coordinate constituent in a parse. Starting from the leaves, we assign each node a polarity based on the polarity of its immediate children following the approach:

if (children contain at least one negative polarity)
  update node polarity as negative
else if (children contain positive polarity)
  update node polarity as positive
else
  update node polarity as neutral
if (node has a child connected with NG (negation) edge)
  reverse node polarity (neutral polarity is switched to negative)

Finally, we assign the polarity of the clause to the targets occurring within the clause.

We evaluate the polarity assignment for the correctly extracted 225 targets in the development and 53 correctly extracted targets in the test corpus against the polarity decisions of a human annotator. Precision (P) reported for a polarity value indicates the

[11] ich habe großes Intresse an
[12] großes Interesse
[13] http://www.ims.uni-stuttgart.de/tcl/SOFTWARE/BitPar.html

proportion of the correctly identified polarity instances for this value to all instances identified with this polarity value by the system. For instance, P for positive polarity is $\frac{correctly\ classified\ positive\ targets}{all\ targets\ classified\ as\ positive}$. Recall (R) of a polarity value is the proportion of the correctly identified polarity instances for this value to the actual number of the polarity instances for this value in the gold standards. For instance, R for positive polarity is $\frac{correctly\ classified\ positive\ targets}{number\ of\ positive\ targets\ in\ gold\ standards}$. Table 3 and Table 2 present the results and the polarity distribution for the correctly extracted targets. The error analysis shows that

Total	Positive	Negative	Neutral
Development (225)	158	23	44
Test (53)	40	3	10

Table 2. Polarity distribution among the correctly extracted targets

Corpus	Positive		Negative		Neutral	
	P	R	P	R	P	R
Development	0.95	0.60	0.73	0.82	0.40	0.90
Test	0.85	0.45	0.50	0.33	0.23	0.70

Table 3. Polarity assignment evaluation for the correctly extracted targets

our opinion lexicon performs satisfactorily at marking the positive sentiment cues most of the time. However, high precision against low recall in positive polarity assignments and the reverse situation in neutral assignments (low precision against high recall) reveal problems with the polarity assignments despite the good coverage of the lexicon.

A major source of errors is the clause-level granularity of analysis. We loose sentiments in the subordinate clauses which refer to targets in the main clauses and vice versa. For instance, we assign neutral polarity to the professional feature `numbers` in the sentence `I'm impressed with the fact that one can explain everything with numbers` as the subordinate clause does not contain any sentiment cue. Furthermore, parsing errors constitute an additional problem in polarity assignments especially when the clause boundaries were not marked correctly in the parse tree. In such cases, we were unable to assign the correct polarity even though we were able to detect the sentiment cue.

We observe relatively good results on the negative polarity assignments in the development corpus compared to the test set. We cannot be very conclusive regarding this on the test corpus due to a small number of the respective negative polarity instances. Again, in negation detection, we see that our approach cannot detect the long distance negation, i.e., the negation spanning the subordinate clauses.

4. Conclusions

We presented an unsupervised lexicon based sentiment analysis component and its intrinsic evaluation in a career recommendation scenario. We extracted *professional preferences* defined as *(target, polarity)* tuples, where *targets* are the terms and phrases facilitating automatic career recommendation and *polarities* are user's preferences regarding the targets. In target extraction, we utilized GermaNet and syntactic patterns over the

results of shallow parsing. The results of *target* extraction show that the lexicon based approach is tied to the coverage and the quality of the lexicon.

We also applied a lexicon based approach to sentiment analysis. The approach assumes that each clause represents an individual unit. Hence, we assigned the polarity of a clause to the *targets* within the clause. The performance of our approach, on one hand, relies on the accuracy of the parser which is sometimes erroneous, and on the other hand on the number of sentences, in which we fail to associate the sentiment cue with the target due to long distance negations or modifications in the subordiante clauses. Nevertheless, unlike statistical methods currently dominating the field of sentiment analysis our approach does not require any training data which is very expensive to obtain for new domains.

During the manual annotation study, we observed that discourse level analysis and coreference resolution play important roles in the correct interpretation of one's preferences. Users tend to express their opinions in a sequence of sentences, where first sentence contains the target, and subsequent sentences contain preferences regarding the target using references to the target. We plan to incorporate these aspects in our future work.

Acknowledgements

This work was supported by the German Research Foundation (DFG) under the grant *GU 798/1-2*, *Semantic Information Retrieval from Texts in the Example Domain Electronic Career Guidance*, and by the Volkswagen Foundation as part of the Lichtenberg-Professorship Program under grant No. I/82806.

References

[1] I. Gurevych, C. Müller, and T. Zesch, "What to be? - electronic career guidance based on semantic relatedness," in *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, (Prague, Czech Republic), pp. 1032–1039, June 2007.

[2] J. Wiebe, T. Wilson, and C. Cardie, "Annotating expressions of opinions and emotions in language," *Language Resources and Evaluation*, vol. 39, pp. 165–210, 2005.

[3] V. Cvoro, "Recognition of emotional preferences for professional features," Master's thesis, Ruprecht-Karls University, Department of Computational Linguistics, Heidelberg, August 2005.

[4] C. Kunze, *Lexikalisch-semantische Wortnetze*, ch. Computerlinguistik und Sprachtechnologie, pp. 423–431. Spektrum Akademischer Verlag, 2004.

[5] H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *Proceedings of Conference on New Methods in Language Processing*, (Manchester, UK), pp. 44–49, September 1994.