Automatic Expression Attribution in Non-Standard Texts

Andreas Müller, Nils Reiter

Institut für Maschinelle Sprachverarbeitung, Stuttgart University

{andreas.mueller,nils.reiter}@ims.uni-stuttgart.de

ABSTRACT

Recognizing and attributing expressions from other authors in texts is an important task for the automatic analysis of scientific texts. Expressions are a more general type of utterance than quotations because they are not limited to what an author explicitly said or wrote, but can consist of references to any type of content expressed by an author. In this paper we describe an ongoing effort to detect expressions of an author in poetics, a genre of literary scholarly writing about poetry. Our method builds on shallow linguistic processing of texts and the availability of structured (linked) data. The method detects sentences containing such expressions and links the detected utterers to their entries in DBPedia. We report initial results showing that a simple technique that makes use of semantic knowledge can achieve reasonable performance for the detection of attributed expressions. The results of the detection are used to (i) highlight expressions, thus allowing for quicker browsing through large quantities of texts and (ii) link the utterers to their respective DBPedia entry, in turn enabling the creation of semantically enriched representations of the documents.

KEYWORDS: digital humanities, linked data, DBPedia, shallow linguistic processing, expression attribution.

1 Introduction

In this paper, we present an approach to the automatic detection of attributed expressions in texts, which is tailored to the needs of digital humanities projects. Our approach makes use of semantic information retrieved from existing knowledge bases and minimizes dependency on linguistic annotations.

We aim at detecting all kinds of attributed expressions and – in contrast to related work – not only direct/indirect speech or quotations. While direct/indirect speech is used to reproduce utterances more or less exactly (1a), other kinds of attributed expressions may also summarize or paraphrase utterances (1b).

(1) a. Goethe hat ihm daraufhin zur Antwort gegeben: [...]
(To this, Goethe answered him: [...])

b. Es ist, wie Husserl gezeigt hat, widersinnig zu sagen, sie können schwanken.
(It is, as Husserl showed, paradoxical to say they could vary)

These kinds of attributed expressions are often used in scientific writing, where the reporting of previous work is of great importance. For our study, we are employing a corpus of scientific writing that is central to literary analysis: A collection of German poetics, written between the 18th and 20th century. This corpus was manually chosen from a larger corpus of 1000 German poetics analyzed by (Richter, 2010). The texts describe important aspects of literature, such as definitions of central terms like "poetry" or a specific type of poetry like "dramatic poetry". In this scenario, we are aiming at providing researchers from literary analysis a quick and easy way to inspect references and relations between the texts. Concretely, we seek to identify triples in the form of `(person, verb-cue, sentence-id)`, so that text segments can be linked to existing knowledge bases or visualized in a number of ways.

We are refraining from basing our approach on linguistic structures, because the corpus contains texts in (multiple) language varieties and thus accurate linguistic processing of the texts is difficult. Instead, we are exploiting the fact that structured semantic information already exists for our domain.

2 Related Work

In the area of reported speech detection and attribution, which is closely related to our task, two general approaches can be discerned. *Rule-based systems* extract quoted speech based on verb lexicons and syntactic or lexical patterns. The system presented by (Krestel et al., 2008), for instance, achieves a performance of 99% precision and 74% recall for the detection of quotation spans on the Wall Street Journal, using a list of 53 reporting verbs and 6 lexical rules. *Supervised systems* are trained on manually labeled data and employ various features to identify quoted speech. (O'Keefe et al., 2012) focus on speech attribution using a sequence labeling approach with surface-oriented features. (Pareti et al., 2013) describes two approaches for quotation extraction, one based on tokens (sequence labeling) and one based on constituents created by a parser (classification). Pareti et al. report that the token-based approach achieves the highest performance, averaged over all kinds of quotations. Most recently, (Almeida et al., 2014) have described a joint approach to quotation attribution and coreference resolution, using a logic program and dual decomposition. For both tasks, lexical and shallow semantic features (e.g., gender) are used. The authors report that the joint approach outperforms pipelined

variants. (Brunner, 2013) is (to our knowledge) the only publication in this area using German (narrative) texts, with the aim of identifying sentences containing speech, thought and writing representation. The author experimented with rules based on lexical information and statistical methods, using POS and surface features. He concludes that rule-based systems work better on strictly defined cases like direct and indirect speech, while machine-learning systems achieve better results on freer cases.

To summarize: Existing approaches focus on various forms of reported speech and employ syntactic and shallow features. Semantic knowledge about potential utterers has not been used in previous work.

3 Method

Our method for extracting triples of `(author, verb-cue, sentence-id)` relies on the following linguistic pre-processing steps: Sentence splitting, tokenization, lemmatization, morphological tagging (using a component from the pipeline described in (Björkelund et al., 2010)) and NER (Faruqui and Padó, 2010). After linguistic pre-processing we apply the following steps: (i) classification of persons, (ii) detection of quotations and (iii) Extraction of relations.

Person Classification In order to identify authors (which are potential sources of expressions) and to distinguish character names from author names, we utilize the DBPedia person data set[1]. If a name is contained in the DBPedia data set, we consider it as a candidate for being an author. So for every person name *A* we find in a poetic, we extract a list of possible matches from the DBPedia data set. If multiple instances are found in the DBPedia data set, we consider each of them as a possible match and use the following two filtering steps to decide which person *A* references.

The first step is to remove all persons which are born after the author of the poetic died from the list of possible matches. The birthdate of a person can be extracted from DBPedia, and we obviously know when the author of the poetic we are investigating died.

The second step aims at filtering according to occupations/professions. Most entries for a person in the DBPedia data set contain a short description of the person. This description frequently mentions the profession of the person. Having a predefined list of "valid" professions (defined by the domain expert participating in our project) allows us to scan the description of a person for these professions. If the description of a person does not contain an occupation from the list, the person is removed from the list of possible matches. If the description contains an occupation from the list, we assign a rank to the person depending on which priority the occupation has. Occupations are prioritized based on how plausible the domain expert participating in our project thought it was that a person having that occupation is mentioned as an author in a poetic. For example, "poet" is an occupation with the highest priority, because it is very likely that the works of poets are referenced in a poetic.

The remaining members of the list of possible matches are then sorted by priority of their occupation. If there is one member with the highest priority, this member is chosen as the match. If there is more than one such member we extract all those members as possible matches. If there is no such member (meaning that the list is empty), we don't consider the person name *A* to be a name which refers to a person who is an author of an expression.

[1] http://wiki.dbpedia.org/Downloads39, Dataset Persondata for German

Positive	Full	60.7 %
	Partial	14.3 %
Negative		25 %

Table 1: Evaluation results

This approach to named entity disambiguation is tailored to the specific domain we are working in, where a relatively small set of professions is relevant, and those are mentioned in the DBpedia descriptions. Since this is not necessarily the case on other domains and tasks, we are planning to extend it to a more general approach.

Quotation Detection As has already been discussed in previous work, quotation marks are a strong indicator of quotations (Brunner, 2013). In many texts, however, quotation marks are used ambiguously for quotations, titles or to indicate emphasis or irony.

In a first step, we use four lexical patterns to recognize text between quotation marks as titles. For example, one of the patterns is `name (genitiv) text_within_quotation_marks` (e.g. "Goethes 'Faust' "). We observe that those patterns are very precise, so the next step only looks at text between quotation marks which was not classified by the first step, in order to increase coverage. In a second step, we identify a text between quotation marks as a quotation if it is not recognized as a title and if there is a person name in the same sentence. For recognizing titles, we use a list of 86,253 titles extracted from the publicly available TextGrid corpus (Hedges et al., 2013). A text between quotation marks is then a title if the text is a member of this list of titles. To compensate for common words being emphasized we compare the frequency of an n-gram between quotes with how often that n-gram appears outside of quotation marks. If it appears often outside quotation marks chances are it is a common word which is emphasized rather than the title of a work or a quotation. Both the recognition based on the extracted titles and the frequency-based filtering are only applied to text within quotation marks which was not classified by the first step (the recognition based on lexical patterns).

Relation Extraction For each sentence, we extract an expression relation if either a) a quote and a person name are contained in the sentence or b) an author name and a verb from a verb list are contained in a sentence. The verb list we use consists of manual translations of the direct troponyms of the verb "express" in the sense of "verbalize, verbalise, utter, give tongue to" as found in WordNet. This results in a list of 23 verbs. We used WordNet because it contains a lot of troponyms for the verb "express", which seems to express the expression relation most closely.

4 Evaluation

We evaluated our technique by extracting triples from the poetic "Grundbegriffe der Poetik" (Staiger, 1946). The system identified 56 instances of attributed expressions within the text. We then manually classified the instances into three classes, in order to gain insight into the behavior of the algorithm and to guide future work. If the sentence contained an attributed expression and the utterer was identified correctly, we considered the instance to be annotated fully correct. If the sentence contained an attributed expression but the utterer was not correctly identified, we considered the item to be partially correct. All other instances were considered an error. The authors of this paper annotated these classes in parallel, with an initial F_1-agreement of 0.67. Differences have been adjudicated after discussion with a domain expert.

Table 1 shows the distribution of the extracted instances over these classes. As we can see, the majority of the items (75%) are correctly identified as expressions. Many of those are also linked to the correct utterer.

Error Analysis We made a detailed investigation of the remaining errors the system made and identified one major error source. The majority of the erroneously identified sentences do contain an expression, but not one of an author within our domain of interest (e.g., a fictional character).

5 Conclusions

We showed a simple technique for extracting (author, express, expression) triples from poetics. Our method uses semantic knowledge about persons contained in the DBPedia database and domain-dependent resources to address the special challenges poetics present. Despite those challenges our technique achieves moderate to good results. Further, it is relatively easy to implement our technique for different languages because it is not based on syntactic parsing.

References

Almeida, M. S. C., Almeida, M. B., and Martins, A. F. T. (2014). A joint model for quotation attribution and coreference resolution. In *Proceedings of EACL*, pages 39–48, Gothenburg, Sweden.

Björkelund, A., Bohnet, B., Hafdell, L., and Nugues, P. (2010). A high-performance syntactic and semantic dependency parser. In *Proceedings of the 23rd International Conference on Computational Linguistics: Demonstrations*, COLING '10, pages 33–36, Stroudsburg, PA, USA. Association for Computational Linguistics.

Brunner, A. (2013). Automatic recognition of speech, thought, and writing representation in german narrative texts. *Literary and Linguistic Computing*, 28(4):563–575.

Faruqui, M. and Padó, S. (2010). Training and evaluating a german named entity recognizer with semantic generalization. In *Proceedings of KONVENS*, Saarbrücken, Germany.

Hedges, M., Neuroth, H., Smith, K. M., Blanke, T., Romary, L., Küster, M., and Illingworth, M. (2013). Textgrid, textvre, and dariah: Sustainability of infrastructures for textual scholarship. *Journal of the Text Encoding Initiative [Online]*.

Krestel, R., Bergler, S., and Witte, R. (2008). Minding the source: Automatic tagging of reported speech in newspaper articles. In *Proceedings of LREC*, Marrakech, Morocco.

O'Keefe, T., Pareti, S., Curran, J. R., Koprinska, I., and Honnibal, M. (2012). A sequence labelling approach to quote attribution. In *Proceedings of EMNLP-CoNNL*, pages 790–799, Jeju Island, Korea.

Pareti, S., O'Keefe, T., Konstas, I., Curran, J. R., and Koprinska, I. (2013). Automatically detecting and attributing indirect quotations. In *Proceedings of EMNLP*, pages 989–999, Seattle, Washington, USA.

Richter, S. (2010). *A History of Poetics: German Scholarly Aesthetics and Poetics in International Context, 1770-1960*. De Gruyter, Berlin, New York: de Gruyter.

Staiger, E. (1946). *Grundbegriffe der Poetik*. Atlantis Verlag AG, Zürich.