Named Entity Linking in #Tweets with KEA

Jörg Waitelonis, Harald Sack
Hasso-Plattner-Institute
Prof.-Dr.-Helmert Str. 2-3, 14482 Potsdam, Germany
{joerg.waitelonis|harald.sack}@hpi.de

ABSTRACT

This paper presents the KEA system at the #Microposts 2016 NEEL Challenge. Its task is to recognize and type mentions from English microposts and link them to their corresponding entries in DBpedia. For this task, we have adapted our Named Entity Disambiguation tool originally designed for natural language text to the special requirements of noisy, terse, and poorly worded tweets containing special functional terms and language.

Keywords

named entity linking, disambiguation, microposts

1. INTRODUCTION

Microposts have become a highly popular medium to share facts, opinions or emotions. They provide an invaluable real-time resource of data, ready to be mined for training predictive models. However, the effectiveness of existing analysis tools faces critical challenges when applied to microposts. In fact it is seriously compromised, since Twitter messages often are noisy, terse, poorly worded and posted in many different languages. They contain special functional expressions, such as e. g. usernames, hashtags, retweets, abbreviations, and cyber-slang [2]. Moreover, Twitter being the most popular micropost service follows a streaming paradigm imposing that entities must be recognized in real-time.

In this paper, we describe our approach to address the #Micropost 2016 NEEL challenge with the adaptation of an existing Named Entity Disambiguation system – KEA – originally designed for the processing of natural language texts to the special challenges imposed by microposts.

KEA originally implements a dictionary and knowledge-based approach of word sense disambiguation, i.e. co-occurrence analysis based on articles of the English Wikipedia are combined with a link-graph analysis on the Wikipedia hyperlink graph and the DBpedia knowledge base. The basic principles of the KEA named entity linking are summarized in [3]. A comparison of KEA and other state-of-the-art named entity linking systems is provided in [5].

In the subsequent sections, KEA will be introduced in more detail, followed by adaptions made especially for the NEEL challenge, and our achieved results.

2. THE KEA APPROACH

To address the tasks of the #Micropost 2016 NEEL challenge, we have adapted our NEL approach KEA. It is originally configured to be applied on natural language text and combinations of textual metadata from heterogeneous sources such as e. g. metadata generated by automated multimedia analysis or user provided metadata, such as e. g. tags, comments, and discussions. All this metadata can be of different provenience, reliability, trustworthiness, as well as level of abstraction. KEA has been successfully deployed within the GERBIL General Entity Annotation Benchmark Framework [5].

KEA uses DBpedia as a reference knowledge base for entity linking and basically follows the five-stage approach depicted in Fig. 1.

2.1 Preprocessing

The incoming text is processed by the following linguistic pipeline. The Stanford Log-linear tagger[4] as well as Stanford Named Entity Recognizer[1] (NER) are applied to determine part-of-speech as well as named entity types. Next, an ASCII folding filter converts alphabetic, numeric, and symbolic Unicode characters, which are not in the the "Basic Latin" Unicode block into their ASCII equivalents, e.g. "Ole Rømer" is transformed to "Ole Romer". Tokenization is performed on non-characters except special characters joining compound words, such as, e.g. "-".

The resulting list of tokens is fed into a shingle filter to construct token n-grams from the token stream. For example, the sentence "please divide this sentence into shingles" might be tokenized into 2-shingles "please divide", "divide this", "this sentence", "sentence into", and "into shingles". Usually, 3-shingles are created as a default. In the case of a proper noun recognized by the NER at most 5-shingles are created with the ±2 surrounding tokens. This extension enables to map also longer compound proper names such as, e.g., "John F. Kennedy Airport" which cannot be mapped correctly otherwise with a 3-shingle configuration. The to-

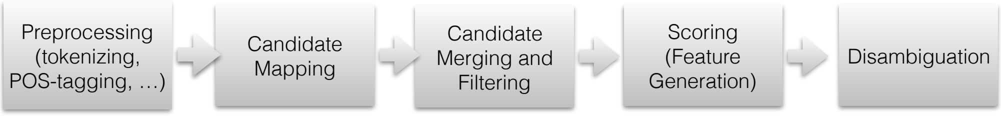
Preprocessing (tokenizing, POS-tagging, ...)
Candidate Mapping
Candidate Merging and Filtering
Scoring (Feature Generation)
Disambiguation

Figure 1: The overall NEL process.

ken stream now contains tokens with sole words, but also tokens with 'shingled' words.

2.2 Candidate Mapping

Every token is mapped to a gazetteer, which has been compiled from DBpedia entities' labels, redirect labels, and disambiguation labels being mapped to their appropriate DBpedia entities. Since the originally used gazetteer in KEA is based on DBpedia 3.9, entities and labels from the DBpedia 2015-04 dataset are added for the NEEL challenge. Labels are indexed lowercase and finally mapped to the tokens resulting in a list of potential entity candidates for each token. The mapping is obtained by exact matches only. A normalization of simple plural forms is applied beforehand. Hence, for each token of the token stream a set of potential entity candidates is determined.

2.3 Candidate Merging and Filtering

To resolve possible overlaps of tokens successfully, longer tokens, which are mapped successfully, are preferred over shorter ones. Since longer tokens contain more descriptive terms, they are considered to be more specific. This means, for example, that "new york city" is preferred over "new york" and "york city". Furthermore, tokens are discarded, if they do not contain nouns or contain sole stopwords, i. e. token "the times" will not be discarded, because it contains the noun "times".

2.4 Scoring (Feature Generation)

For every entity candidate, features are determined via a pipeline of analysis components (scorers). These components asses different characteristics how well a candidate entity fits to the given input text, which is considered as the context. We distinguish between local and context-related features. Local features only consider the candidate as well as the tokens properties. For example, consider the text "Armstrong landed on earth's satellite": For a candidate w.l.o.g "dbp:Neil_Armstrong" of the possible candidate list of the token 'Armstrong' certain features can be determined, as e. g. string-distance between the candidate labels and the token (respectively the surface form), the candidates link graph popularity, its DBpedia type, the provenance of the label, the surface form matches best (e. g. main label, or redirect label), or the level of ambiguity of the token (e. g. approximated by the number of candidates).

Context-features assess the relation of a candidate entity to the other candidates within the given context, e. g. direct links to other context candidates in the DBpedia link graph, co-occurrence of the other tokens' surface forms in the corresponding Wikipedia article of the candidate under consideration, co-references in Wikipedia articles, as well as further graph based features of the link graph induced by all candidates of the context (context graph). This includes for example, graph distance measurements, connected component analysis, or centrality and density observations.

Overall, after this processing step, every candidate gets a list of scores assigned being determined via several of the mentioned methods. Theses lists of scores are considered as the candidates' feature vectors, expressing how well a candidate entity fits to the given context.

2.5 Disambiguation

Since all scores of the analyzed features have a positive but unlimited value range, a linear feature scaling is applied to standardize the ranges between 0.0 and 1.0. Different approaches ranging from statistical analysis to machine learning techniques can be envisaged to decide which candidate is chosen as the winner for a token. The most basic approach considers the weighted sum of the scores as a confidence score, whereas the weights are optimized via grid search on a given development or training dataset. The confidence score is cut-off by a empirically optimized threshold, to decide, if a candidate entity is to be selected as the assumed correct result.

3. ADAPTATIONS TO THE NEEL CHALLENGE

To be applicable also for microposts as in the NEEL challenge, the KEA processing has been adapted in two ways. We distinguish between modifications made especially for the general domain of "microposts/tweets" and modifications resulting from the observation of the provided training data set.

3.1 Adaptations to the Domain

For the NEEL challenge, we have utilized characteristic tweet information by excluding "@" and "#" from the tokenization to later identify twitter user names and hash tags properly. With respect to the provided NEEL challenge guidelines of annotations, the filter is extended to restrict the system to tokens containing singular and plural proper nouns, user names, as well as hashtags only. The stopword list is extended with twitter specific functional terms (e.g. "RT", "MT", etc.) to be ignored in further processing. KEA is configured to consider a single micropost (tweet) as the given context for disambiguation. Furthermore, the threshold of the achieved confidence score is used to cut-off uncertain candidates resulting in NIL annotations. Tokens identified as user name or hashtag which cannot successfully be mapped to candidate entities are also annotated with NIL.

3.2 Adaptations to the Training Set

From the provided training dataset all surface forms have been extracted to extend the gazetteer for candidate mapping. We have optimized the scorer weights as well as the overall threshold according to the results achieved for the training and development datasets. Furthermore, the stopword list has been extended according to the achieved results from the training and development datasets, i.e. terms constantly mapped wrongly because they have not been annotated in the datasets such as weekdays and months.

3.3 Types

Since KEA did not support the required annotation with types out of the box, a simple extension of the original framework has been implemented. For a disambiguated mapped entity, type annotations are determined simply via lookup in the DBpedia instance types dataset. For NIL annotations, where no entity could be determined, the according NER type, if available, has been chosen.

4. EXPERIMENTS AND RESULTS

For the #Microposts 2016 NEEL challenge we have first analyzed the provided development dataset without the above described adaptions to obtain a baseline (cf. Table 1), and then again with the NEEL challenge modifications (cf. Table 2).

Table 1: Results for the NEEL2016 development data set (baseline, without modifications)

Measure	Prec.	Recall	F_1score
strong link match	0.399	0.490	0.440
strong typed mention match	0.232	0.213	0.222
mention ceaf	0.611	0.562	0.586

Table 2: Results for the NEEL2016 development data set after adaptions and optimization

Measure	Prec.	Recall	F_1score
strong link match	0.667	0.862	0.752
strong typed mention match	0.572	0.660	0.613
mention ceaf	0.744	0.858	0.797

According to our expectations, the special adaptations for the NEEL challenge have resulted in significantly better results compared to the original tool configuration. A closer inspection of the achieved mappings has shown that KEA was able to find correct mappings to entities which are not provided in the NEEL ground truth, e. g.:

```
#wcyb -> dbp:WCYB-TV
#WSJ ->  dbp:The_Wall_Street_Journal
#NSC -> dbp:National_Security_Council
#kyloren -> dpb:Kylo_Ren
```

Compared to the training data ground truth, the KEA system tends to detect mentions overeagerly, i. e. the system produces more extra annotations than missing annotations, which results in a loss of precision. Many of KEA's extra annotations are common nouns such as affirmative action, astronaut, petition, signature, mosque, emoji, enemy.

5. CONCLUSION & FUTURE WORK

For the task of NEL on microposts, it is a challenge to maintain the topicality of the underlying knowledge base. New hash-tags, neologisms, as well as cyber-slang are rather difficult to resolve correctly in an automated way because they are not present in the dictionaries. To cope with this situation, one possibility would be to include a live analysis of the Wikipedia update stream to extend or prioritize the used dictionary of surface forms as well as the underlying link graph.

From our observations, a significant part of the achieved improvements results from the fact that training sets as well as test sets cover the identical domains (i.e. Star Wars and Donald Trump). Hence, the extension of the dictionary with surface forms of the training dataset seems to be very effective. The conclusion is, that a domain adaption for a given general purpose system might lead to significantly better results. Even if this sounds trivial, we did not expect an improvement of c. 40% in f-measure.

Unfortunately, many documents of the training data set (1951 out of 6024) do not have any annotations at all. Therefore, we are looking forward to future NEEL challenges with more complete ground truth datasets.

6. REFERENCES

[1] J. R. Finkel, T. Grenager, and C. D. Manning. Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, pages 363–370, 2005.

[2] B. Han and T. Baldwin. Lexical Normalisation of Short Text Messages: Makn Sens a #Twitter. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, HLT '11, pages 368–378, Stroudsburg, PA, USA, 2011.

[3] H. Sack. *Business Information Systems Workshops: BIS 2015 International Workshops, Poznań, Poland, June 24-26, 2015, Revised Papers*, chapter The Journey is the Reward - Towards New Paradigms in Web Search, pages 15–26. Springer International Publishing, Cham, 2015.

[4] K. Toutanova and C. D. Manning. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-speech Tagger. In *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics - Volume 13*, EMNLP '00, pages 63–70, Stroudsburg, PA, USA, 2000. Association for Computational Linguistics.

[5] R. Usbeck, M. Röder, A.-C. Ngonga Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – General Entity Annotation Benchmark Framework. In *Proceedings of the 24th International Conference on World Wide Web (WWW15)*, pages 1133–1143. ACM, USA, 2015.