A Ranking Approach to Target Detection for Automatic Link Generation

Jiyin He and Maarten de Rijke
ISLA, University of Amsterdam, The Netherlands
{j.he, derijke}@uva.nl

ABSTRACT

We focus on the task of target detection in automatic link generation with Wikipedia, i.e., given an N-gram in a snippet of text, find the relevant Wikipedia concepts that explain or provide background knowledge for it. We formulate the task as a ranking problem and investigate the effectiveness of learning to rank approaches and of the features that we use to rank the target concepts for a given N-gram. Our experiments show that learning to rank approaches outperform traditional binary classification approaches. Also, our proposed features are effective both in binary classification and learning to rank settings.

Categories and Subject Descriptors: H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software

General Terms: Algorithms, Experimentation

Keywords: Link generation, disambiguation, learning to rank, Wikipedia

1. INTRODUCTION

Annotating text with human defined concepts from ontologies or knowledge bases is useful in providing background knowledge that helps users to understand difficult concepts as well as to capture the meaning of ambiguous words or phrases. Recently, several approaches have been proposed for annotating free text with human defined concepts from Wikipedia or similar knowledge bases in the context of automatic link generation. Here we focus on a sub-task of the automatic link generation problem, namely, the target detection task. That is, given a N-gram in a piece of text, find the related concepts in Wikipedia, i.e., Wikipedia pages. Mihalcea and Csomai [6] formulate the target detection task as a classification problem and propose several linguistic features for learning to link the phrases in free texts to Wikipedia pages. Milne and Witten [5] improve the classification performance by utilizing contextual features, i.e., the surrounding "non-ambiguous" words or phrases, of a given N-gram. Meij et al. [4] study the problem of semantic query suggestion, where each query is linked to a list of concepts from DBpedia, ranked by their relevance to the query. Although presented as a ranking problem, they use binary classification to rank the related concepts. INEX (the INitiative for the Evaluation of XML retrieval) has launched the Link-the-Wiki task, which defines target detection as a ranking problem, as at most 5 target concepts can be returned for an anchor text; various heuristics as well as retrieval-based methods have been proposed [7].

We investigate the effectiveness of learning to rank approaches for target detection. By its very nature, the target detection task is a ranking problem. First, the recommended concepts from the knowledge base may give background knowledge at different granularity levels. For example, given the N-gram "dublin core," while the Wikipedia page on "dublin core" is an obvious choice, the concept "ontology" may also be interesting for a reader unfamiliar with the topic. Second, when context is short, it is usually difficult to be certain about the meaning of an ambiguous word or phrase. Listing possible concepts ordered by decreasing relevance thus effectively avoids missing correct answers while trying to provide the correct answers as early as possible in a ranked list. We explore features that do not rely heavily on the "non-ambiguous" context of a given N-gram (as proposed in [5]), which makes our method easily applicable to short texts such as queries submitted to a search engine.

We experiment with the INEX 2008 Wikipedia collection. The performance achieved by the proposed features is comparable to the state-of-the-art when evaluated with binary classification metrics. However, when measured with ranking-based metrics, our learning to rank approaches using these features significantly outperform binary classification.

2. METHOD

Features We briefly introduce our notation and proposed features. Given a N-gram, we refer to the text snippet (e.g., an article, a paragraph, etc) in which it is contained as a *topic* and a Wikipedia page to which it links (or should link) as a *target*. We consider three types of feature for our learning experiments, namely, *N-gram-target* features, *target* features and *topic-target* features.

N-gram-target features. The N-gram-target features describe how well an N-gram ng and a candidate target $ctar$ are related. Three features are explored in this category: (i) TitleMatch: in Wikipedia, titles usually denote the key concept on which a page focuses and therefore the match between a ng and the title of a $ctar$ indicates the relatedness of the two; we use three values for this feature: 0 (no match), 1 (partial match) and 2 (exact match). (ii) Link Evidence: existing links among Wikipedia pages are effective indicators of how likely ng and $ctar$ are linked, which we measure with the following two scores: *RatioLink* and *RatioAnchor*, calculated as $RatioLink(ng, ctar) = |link(ng, ctar)| \cdot |inlink(ctar)|^{-1}$ and $RatioAnchor(ng, ctar) = |link(ng, ctar)| \cdot |ng \in A|^{-1}$, where $|link(ng, ctar)|$ is the number of times ng and $ctar$ are linked, $|inlink(ctar)|$ is the number of times $ctar$ is linked by some N-grams and $|ng \in A|$ is the number of times ng is annotated as an anchor text and linked to some concepts in Wikipedia. (iii) Retrieval scores: we use ng as query and compute retrieval scores against $ctar$ as a measure of relatedness of the two. BM25 and the Markov Random Field model (MRF) [1] are used.

Target features. The target features are indicators of how likely

SIGIR'10, July 19–23, 2010, Geneva, Switzerland.
ACM 978-1-60558-896-4/10/07.

$ctar$ alone would be linked with some N-grams in Wikipedia. Four features are used, namely, (i) number of inlinks, (ii) outlinks and (iii) Wikipedia categories associated with $ctar$, which presumably indicate the "popularity" of $ctar$, and (iv) *generality* as proposed in [5], i.e., the level of $ctar$ in the Wikipedia category hierarchy.

Topic-Target features. Features of this type describe the relatedness between the context of ng, i.e., the topic t and $ctar$. Two features are considered: (i) cosine similarity between t and $ctar$, and (ii) retrieval score using title of $ctar$ as query and t as target document, where BM25 is used as the retrieval model.

Learning to Rank the Target Concepts We employ two learning to rank approaches, namely Ranking SVM [2] and AdaRank [3]. To construct the training set, we use anchor texts in the Wikipedia collection, which ensures that there exists a manually linked target concept. Each instance is an $(ng, ctar)$ pair. Relevance judgements are derived from the manual annotations, i.e., for a ng, the manually linked target concept is judged as relevant, while other candidate target concepts are judged as non-relevant. Although this is a binary judgement, the learning algorithms learn the preference relation between the positive and negative examples. There exist various ways to collect candidate target pages for a given ng, e.g., using ng as a query and retrieving a list of potentially relevant Wikipedia pages. Here we use all the pages that have been linked to the ng at least once in Wikipedia as candidate target concepts. Also, we use various binary classifiers as baseline approaches. Since the binary classifiers output binary decisions, a logistic regression model is fit to the output so as to generate probability distributions for the binary decisions such that candidate target concepts are ranked according to their probabilities of being a positive example.

3. EXPERIMENTS AND RESULTS

For our experimental evaluation, we construct the training, validation and test set from the INEX 2008 Wikipedia collection. We randomly sample 500 pages for training, 100 pages for validation, i.e., to tune model parameters, and 50 pages for testing. The training set contains 11,112 anchor texts, which result in 170,102 instances; the validation set contains 9,365 anchors texts and 106,051 instances; the test set contains 3,452 anchor texts and 33,259 instances. We use existing Wikipedia links as ground truth. Following Meij et al. [4], three binary classifiers are used as baselines: NaiveBayes, SVM with linear kernel and J48, a decision tree type classifier. We use WEKA[1] for binary classification and SVMLight for RankingSVM.[2] We use MAP, P@1 and P@5 to measure the ranking performance of the binary classifiers as well as the learning to rank approaches.

Table 1 shows the results. P@1 measures how good the first result in the ranked list is, which is important for the target detection task, as ideally we would expect that the first answer is a correct one while the remaining concepts in the ranked list supply complementary material. P@5 measures the unranked early precision. We see that all methods except NaiveBayes have a similar performance in terms of P@5. This suggests that the difference of the rankings generated by different algorithms are limited to the very top of the ranked list, e.g., within the top 5. In general, learning to rank approaches outperform binary classification approaches. AdaRank significantly outperforms all binary classifiers, on all measures.

In addition, Table 2 shows the binary classification results with our proposed features. J48 achieves best performance, which is comparable to the performance of binary classification approches in the literature. In [6] the best performance (F-measure 0.88) was achieved by applying a NaiveBayes classifier with linguistic features. Milne and Witten [5] report an F-measure of 0.96, achieved using a C4.5 classifier with a set of context-based features. The main difference between our features and those proposed in [5] is that our features do not rely heavily on the context "non-ambiguous" concepts, which allows our method to be successfully applied to short texts with limited context.

Method	MAP	p@1	p@5
Ranking SVM	0.9502	0.9235	0.1970
AdaRank	**0.9629**	**0.9395**	**0.1980**
SVM-class	0.9476$^{\triangledown\blacktriangledown}$	0.9180$^{\triangledown\blacktriangledown}$	0.1970$^{\blacktriangledown}$
J48	0.9496$^{\blacktriangledown}$	0.9218$^{\triangledown\blacktriangledown}$	0.1968$^{\blacktriangledown}$
NaiveBayes	0.9200$^{\triangledown\blacktriangledown}$	0.8699$^{\triangledown\blacktriangledown}$	0.1962$^{\triangledown\blacktriangledown}$

Table 1: Performance of learning to rank approaches compared to binary classification approaches. $^{\triangledown}$ ($^{\blacktriangledown}$) denotes significant difference as determined using a paired t-test at level 0.05 between binary classification and RankingSVM (AdaRank).

Method	Precision	Recall	F-measure
SVM-class	0.86	0.82	0.84
J48	**0.93**	**0.91**	**0.92**
NaiveBayes	0.63	0.67	0.65

Table 2: Performance of binary classifications on the test set.

4. CONCLUSION

We investigated the effectiveness of learning to rank approaches and a set of underlying features for target detection in automatic link generation. Learning to rank approaches outperform binary classifiers in terms of various ranking metrics with the same set of features. Our features are shown to be useful in both binary classification and learning to rank settings. We leave the analysis of the importance of the features for future work. A natural next step is to extend the binary judgements to multiple relevance levels. Also, our approach to target detection can be naturally applied to many real-world problems such as word sense disambiguations as well as semantic query suggestion with Wikipedia.

Acknowledgements This research was supported by the European Union's ICT Policy Support Programme as part of the Competitiveness and Innovation Framework Programme, CIP ICT-PSP under grant agreement nr 250430, by the DuOMAn project carried out within the STEVIN programme which is funded by the Dutch and Flemish Governments under project nr STE-09-12, and by the Netherlands Organisation for Scientific Research (NWO) under project nrs 612.066.512, 612.061.814, 612.061.815, 640.004.802.

5. REFERENCES

[1] D. Metzler and W. B. Croft. A Markov Random Field Model for Term dependencies. In *SIGIR'05*, 2005.

[2] R. Herbrich, T. Graepel and K. Obermayer. Large margin rank boundaries for ordinal regression. In *Advances in Large Margin Classifiers*, 2000.

[3] J. Xu and H. Li. AdaRank: a boosting algorithm for information retrieval. In *SIGIR'07*, 2007.

[4] E. Meij, M. Bron, L. Hollink, B. Huurnink and M. de Rijke. Learning Semantic Query Suggestions. In *ISWC'09*, 2009.

[5] D. Milne and I. H. Witten. Learning to link with Wikipedia. In *CIKM '08*, 2008.

[6] R. Mihalcea and A. Csomai. Wikify!: linking documents to encyclopedic knowledge. In *CIKM '07*, 2007.

[7] G. Shlomo, J. Kamps and A. Trotman. Advances in Focused Retrieval. In *INEX '08*, 2008.

[1] http://www.cs.waikato.ac.nz/ml/weka/

[2] http://svmlight.joachims.org/