Combining Mention Context and Hyperlinks from Wikipedia for Named Entity Disambiguation

Ander Barrena
IXA NLP Group
UPV/EHU
Donostia, Basque Country
ander.barrena@ehu.eus

Aitor Soroa
IXA NLP Group
UPV/EHU
Donostia, Basque Country
a.soroa@ehu.eus

Eneko Agirre
IXA NLP Group
UPV/EHU
Donostia, Basque Country
e.agirre@ehu.eus

Abstract

Named entity disambiguation is the task of linking entity mentions to their intended referent, as represented in a Knowledge Base, usually derived from Wikipedia. In this paper, we combine local mention context and global hyperlink structure from Wikipedia in a probabilistic framework. Our results show that the two models of context, namely, words in the context and hyperlink pathways to other entities in the context, are complementary. We test our method in eight datasets, improving the state-of-the-art results in five, without any tuning, showing that it is robust to out-of-domain scenarios. When tuning combination weights, we match the best reported results on the widely-used AIDA-CoNLL test-b.

1 Introduction

Linking mentions occurring in documents to a knowledge base is the main goal of Entity Linking or Named Entity Disambiguation (NED). This problem has attracted a great number of papers in the NLP and IR communities, and a large number of techniques, including local context and global inference (Ratinov et al., 2011). We propose to use a probabilistic framework that combines entity popularity, name popularity, local mention context and global hyperlink structure, relying on information in Wikipedia alone. Entity and name popularity are useful disambiguation clues in the absence of any context. The local mention context provides direct clues (in the form of words in context) to disambiguate each mention separately. The hyperlink structure of Wikipedia provides a global coherence measure for all entities mentioned in the same context.

The advantages of our method with respect to other alternatives are as follows: (1) It does not involve a large number of methods and classifier combination. (2) The method learns the parameters directly from Wikipedia so no additional hand-labeled data and training is needed. (3) We combine the global hyperlink structure of Wikipedia with a local bag-of-words probabilistic model in an intuitive and complementary way. (4) The absence of training allows for robust results in out-of-domain scenarios.

The evaluation of NED is fragmented, with several popular shared tasks, such as TAC-KBP[1], ERD[2] or NEEL[3]. Other evaluation datasets include AIDA-CoNLL and KORE50[4], which are very common in NED evaluation. Note that each dataset poses different problems. For instance AIDA-CoNLL is composed of news, and systems need to disambiguate all occurring mentions. TAC includes news and discussion forums, and focuses on a large number of mentions for a handful of challenging strings. KORE50 includes short sentences with very ambiguous mentions. Unfortunately, there is no standard dataset, and many contributions in this area report results in just one or two datasets. We report our results on eight datasets, improving the state-of-the-art results on five.

[1] http://www.nist.gov/tac/2014/KBP/

[2] http://web-ngram.research.microsoft.com/erd2014/

[3] http://www.scc.lancs.ac.uk/microposts2015/challenge/index.html

[4] http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/downloads/

2 Resources

The knowledge used by our Bayesian network comes from Wikipedia. We extract three information resources to perform the disambiguation: a dictionary, textual contexts and a graph.

The dictionary is an association between strings and Wikipedia articles. We construct the dictionary using article titles, redirections, disambiguation pages, and anchor text. If the mention links to a disambiguation page, it is associated with all possible articles the disambiguation page points to. Each association between a string and article is scored with the prior probability, estimated as the number of times that the mention occurs in the anchor text of an article divided by the total number of occurrences of the mention. We choose candidate entities for disambiguation by just assigning all entities linked to the mention in the dictionary.

In addition we build a graph using the Wikipedia link structure, where entities are nodes and edges are anchor links among entities from Wikipedia. We used the third-party dictionary and graph described in (Agirre et al., 2015), which is publicly available[5].

Finally, we extract textual contexts for all the possible candidate entities from a Wikipedia dump. We collect all the anchors including a link to each entity in Wikipedia, and extract a context of 50 words around the anchor link.

3 A Generative Bayesian Network

Given a mention s occurring in context c, our system ranks each of the candidate entities e. Figure 1 shows the dependencies among the different variables. Note that context probability is given by two different resources.

Candidate entities are ranked combining evidences from 4 different probability distributions, which we call entity knowledge $P(e)$, name knowledge $P(s|e)$, context knowledge $P(c_{\text{bow}}|e)$ and graph knowledge $P(c_{\text{grf}}|e)$ respectively.

Entity knowledge $P(e)$ represents the probability of generating entity e, and is estimated as follows:

$$P(e) = \frac{Count(e) + 1}{|M| + N}$$

[5] http://ixa2.si.ehu.eus/ukb/ukb-wiki.tar.bz2

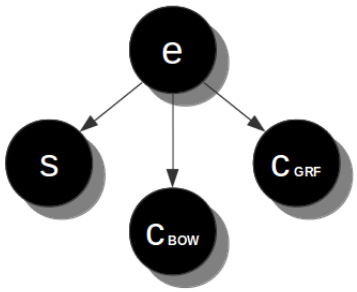

Figure 1: Dependencies among variables in a Bayesian network. The network gives as a result this formula: $P(s, c_{\text{bow}}, c_{\text{grf}}, e) = P(e)P(s|e)P(c_{\text{bow}}|e)P(c_{\text{grf}}|e)$.

where $Count(e)$ describes the entity popularity, e.g., the number of times the entity e is referenced within Wikipedia, $|M|$ is the number of entity mentions and N is the total number of entities in Wikipedia. As can be seen, the estimation is smoothed using the *add-one* method.

Name knowledge $P(s|e)$ represents the probability of generating a particular string s given the entity e, and is estimated as follows:

$$P(s|e) = \theta\frac{Count(e, s)}{Count(e)} + (1 - \theta)\frac{Count(s)}{|M|}$$

where $Count(e, s)$ is the number of times mention s is used to refer entity e and $Count(s)$ is the number of times mention s is used as an anchor. θ parameter is set to 0.9 according to development experiments done in the AIDA-CoNLL development set (also known as AIDA-CoNLL test-a, cf. Section 4).

The context knowledge is modeled in two different ways. In the bag-of-words model, $P(c_{\text{bow}}|e)$ represents the probability of generating context $c = \{w_1, w_2, \ldots, w_n\}$ given the entity e, and is estimated as follows:

$$P(c_{\text{bow}}|e) = P_e(w_1)P_e(w_2)...P_e(w_n)$$

where $P_e(w)$ is estimated as:

$$P_e(w) = \lambda P'_e(w) + (1 - \lambda)P_w(w)$$

$P'_e(w)$ is the maximum likelihood estimation of each word w in the context of e entity. Context words are smoothed by $P_w(w)$ that is the likelihood of words in the whole Wikipedia. λ parameter is set to 0.9 according to development experiments done in AIDA-CoNLL test-a.

The graph knowledge is estimated using personalized Pagerank. We used the probabilities returned

by UKB[6] (Agirre et al., 2015). This software returns $P(e|c_{\text{grf}})$[7] the probability of visiting a candidate entity when performing a random walk on the Wikipedia graph starting in the entity mentions in the context. In order to introduce it in the generative model, we must first convert it to $P(c_{\text{grf}}|e)$. We use Bayes' formula to estimate the probability:

$$P(c_{\text{grf}}|e) = P(e|c_{\text{grf}})P(c_{\text{grf}})/P(e)$$

Finally, the *Full Model* combines all evidences to find the entity that maximizes the following formula:

$$e = \arg\max_e P(s, c_{\text{bow}}, c_{\text{grf}}, e) =$$
$$\arg\max_e P(e)P(s|e)P(c_{\text{bow}}|e)P(c_{\text{grf}}|e)$$

4 Experiments

We tested our algorithms on a wide range of datasets: AIDA-CoNLL test-b (Hoffart et al., 2011), KORE50 (Hoffart et al., 2012) and six TAC-KBP[8] datasets corresponding to six years of the competition (Aida, Kore and Tac hereafter). No corpus was used for training the parameters of the system, apart from Wikipedia, as explained in the previous sections.

We used gold-standard mentions and we evaluated only those mentions linked to a Wikipedia entity (ignoring so-called NIL cases). Depending on the dataset, we used the customary evaluation measure: micro-accuracy (Aida, Kore, Tac09 and Tac10) or Bcubed+ (Tac11, Tac12, Tac13 and Tac14)[9].

Each gold standard uses a different Wikipedia version: 2010 for Aida and Kore, 2008 for Tac. We use the Wikipedia dump from 25-5-2011 to build our resources, as this is close to the versions used at the time. We mapped gold-standard entities to 2011 Wikipedia automatically, using redirects in the 2011 Wikipedia. This mapping could cause a small degradation of our results.

4.1 Results

The top 4 rows in table 1 show the performance of the different combinations among probabilities. The remaining row shows the best results reported to date on those datasets (see caption for details).

The results suggest that each probability contributes to the final score of the *Full Model*, shown on row 4, showing that both context models are complementary between each other[10]. The only exception is Tac13, where the bow model is best.

Our system obtains very good results in all datasets, excelling in Tac09-10-11-12-13, where it beats the state-of-the-art. The figures obtained by the *Full Model* on Aida, Kore and Tac14 are close to the best results. Note that the table shows the results of the system reporting the best values for each dataset, that is, our system is compared not to one single system but to all those systems. For example, (Hoffart et al., 2012) reported lower figures for Kore, 64.58. Regarding the results for TAC-KBP, the full task includes linking to the Knowledge Base and detecting and clustering NIL mentions. In order to make results comparable to those for in Aida and Kore, the table reports the results for mentions which are linked to the Knowledge Base, that is, results where NIL mentions are discarded.

5 Adjusting the model to the data

We experimented with weighting the probabilities to adapt the *Full Model* mentioned above to a specific scenario. For the *Weighted Full Model*, we introduce the α, β, γ and δ parameters[11] as follows:

$$e = \arg\max_e P(s, c_{\text{bow}}, c_{\text{grf}}, e) =$$
$$\arg\max_e P(e)^{\alpha}P(s|e)^{\beta}P(c_{\text{bow}}|e)^{\gamma}P(c_{\text{grf}}|e)^{\delta}$$

Weighting may change the optimal configuration for θ and λ, we thus optimized all parameters on the development set of Aida, yielding $\theta = 0.9$, $\lambda = 0.7$, $\alpha = 0.2$, $\beta = 0.1$, $\gamma = 0.6$ and $\delta = 0.1$ performing a exhaustive grid search. The step size used in this experiment is 0.1. The parameters yielded high results for development, up to 83.48.

Table 2 summarizes the results of the *Weighted Full Model* for Aida, showing that model reaches 84.88 points, a la par to the best micro accuracy reported by (Houlsby and Ciaramita, 2014) and above

[6] http://ixa2.si.ehu.es/ukb/

[7] Note that, contrary to us, the results in (Agirre et al., 2015) multiply the Pagerank probability with the prior.

[8] http://www.nist.gov/tac/publications/index.html

[9] Note that Tac14 results correspond to the so-called Diagnostic Entity Linking task.

[10] The results of our combination involving the UKB software are not comparable to those reported by (Agirre et al., 2015), due to the different formulation of the probability distribution which involves the prior.

[11] $\alpha + \beta + \gamma + \delta = 1$

Test	Aida	Kore	Tac09	Tac10	Tac11	Tac12	Tac13	Tac14
$P(e)P(s\|e)$	67.54	35.42	67.04	76.96	67.83	46.20	66.54	62.01
$P(e)P(s\|e)P(c_{\text{bow}}\|e)$	75.05	60.42	77.19	85.20*	75.55	57.06	**74.56***	71.21
$P(e)P(s\|e)P(c_{\text{grf}}\|e)$	76.83	54.86	79.40*	83.92*	79.75	70.13*	70.21	71.28
$P(e)P(s\|e)P(c_{\text{bow}}\|e)P(c_{\text{grf}}\|e)$	**83.28**	**70.83**	**82.21***	**85.98***	**81.85***	**71.65***	73.99*	**76.48**
Best (state-of-the-art)	84.89	71.50	79.00	80.60	80.10	68.50	71.80	79.60

Table 1: Bold marks the best value among probability combinations, and * those results that overcome the best value reported in the state-of-the-art: (Houlsby and Ciaramita, 2014) for Aida, (Moro et al., 2014) for Kore, (Han and Sun, 2011) for Tac09 and see TAC-KBP proceedings for the rest[8].

Test	Aida
$P(e)P(s\|e)P(c_{\text{bow}}\|e)P(c_{\text{grf}}\|e)$	83.28
$P(e)^{\alpha}P(s\|e)^{\beta}P(c_{\text{bow}}\|e)^{\gamma}P(c_{\text{grf}}\|e)^{\delta}$	84.88
(Moro et al., 2014)	82.10
(Hoffart et al., 2011)	82.54
(Houlsby and Ciaramita, 2014)	84.89

Table 2: Micro accuracy results for Aida introducing the *Weighted Full Model* in row 2.

those reported by (Hoffart et al., 2011; Moro et al., 2014) (respectively, 82.54[12] and 82.10). Unfortunately the parameter distribution seems to depend on the test dataset, as the same parameters failed to improve the results on the other datasets.

6 Related Work

The use of Wikipedia for named entity disambiguation is a common approach in this area. In the related field of Wikification, (Ratinov et al., 2011) introduced the supervised combination of a large number of global and local similarity measures. They learn weights for each of those measures training a supervised classifier on Wikipedia. Our approach is different in that we just combine four intuitive methods, without having to learn weights for them. Unfortunately they don't report results for NED.

(Moro et al., 2014) present a complex graph-based approach for NED and Word Sense Disambiguation which works on BabelNet, a complex combination of several resources including, among others, Wikipedia, WordNet and Wiktionary. Our results are stronger over Aida, but not on the smaller Kore.

(Hoffart et al., 2011) presents a robust method based on entity popularity and similarity measures, which are used to build a mention/entity graph. They include external knowledge from Yago, and train a classifier on the train part of Aida, obtaining results comparable to ours. Given that we do not train on in-domain training corpora, we think our system is more robust.

The use of probabilistic models using Wikipedia for NED was introduced in (Han and Sun, 2011). In this paper, we extend the model with a global model which takes the hyperlink structure of Wikipedia into account.

(Houlsby and Ciaramita, 2014) presents a probabilistic method using topic models, where topics are associated to Wikipedia articles. They present strong results, but they need to initialize the sampler on another NED system, Tagme (Ferragina and Scaiella, 2012). In some sense they also combine the knowledge in the graph with that of a local algorithm (Tagme), so their work is complementary to ours, but in their case the improvement obtained when using the graph is negligible. They only provide results on Aida, and it is thus not possible to compare their robustness with that of our algorithm.

7 Conclusions and future work

Bayesian networks provide a principled method to combine knowledge sources. In this paper we combine popularity, name knowledge and two methods to model context: bag-of-words context, and hyperlink graph. The combination outperforms the

[12]Note that values by (Hoffart et al., 2011) were reported on a subset of Aida. The micro accuracy results reported in our table correspond to the latest best model from the Aida web site: `http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/aida/`.

state-of-the-art in five out of eight datasets, showing the robustness of the system in different domain and dataset types. Our results also show that in all but one dataset the combination outperforms individual models, indicating that bag-or-word context and graph context are complementary. We show that results can be further improved when tuning the weights on in-domain development corpora, matching the best results on the widely-used AIDA-CoNLL test-b.

Given that Bayesian networks can be further extended, we are exploring to introduce additional models of context into a Markov Random Field algorithm. Our current model assumes that the two models of context (bag or words and graph) are independent given e, and we would like to explore alternatives to relax this assumption.

Acknowledgments

We thank the reviewers for their suggestions. This work was partially funded by MINECO (CHIST-ERA READERS project – PCIN-2013-002-C02-01, and SKaTeR project – TIN2012-38584-C06-02), and the European Commission (QTLEAP – FP7-ICT-2013.4.1-610516). The IXA group is funded by the Basque Government (A type Research Group). Ander Barrena enjoys an PhD scholarship from the Basque Government.

References

E. Agirre, A. Barrena, and A. Soroa. 2015. Studying the Wikipedia Hyperlink Graph for Relatedness and Disambiguation. *ArXiv e-prints*, March.

Paolo Ferragina and Ugo Scaiella. 2012. Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software*, 29(1):70–75.

X. Han and L. Sun. 2011. A Generative Entity-mention Model for Linking Entities with Knowledge Base. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 945–954.

J. Hoffart, M.A. Yosef, I. Bordino, H. Fürstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. 2011. Robust Disambiguation of Named Entities in Text. In *Conference on Empirical Methods in Natural Language Processing, Edinburgh, Scotland, United Kingdom 2011*, pages 782–792.

Johannes Hoffart, Stephan Seufert, Dat Ba Nguyen, Martin Theobald, and Gerhard Weikum. 2012. Kore: Keyphrase overlap relatedness for entity disambiguation. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, page 545554.

Neil Houlsby and Massimiliano Ciaramita. 2014. A scalable gibbs sampler for probabilistic entity linking. In Maarten de Rijke, Tom Kenter, ArjenP. de Vries, ChengXiang Zhai, Franciska de Jong, Kira Radinsky, and Katja Hofmann, editors, *Advances in Information Retrieval*, volume 8416 of *Lecture Notes in Computer Science*, pages 335–346. Springer International Publishing.

Andrea Moro, Alessandro Raganato, and Roberto Navigli. 2014. Entity linking meets word sense disambiguation: a unied approach. *Transactions of the Association of Computational Linguistics*, 2:231–244, May.

L.A. Ratinov, D. Roth, D. Downey, and M. Anderson. 2011. Local and Global Algorithms for Disambiguation to Wikipedia. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, pages 1375–1384. The Association for Computer Linguistics.