Disambiguation of entities in MEDLINE abstracts by combining MeSH terms with knowledge

Amy Siu, Patrick Ernst, Gerhard Weikum
Max Planck Institute for Informatics
66123 Saarbrücken, Germany
{siu, pernst, weikum}@mpi-inf.mpg.de

Abstract

Entity disambiguation in the biomedical domain is an essential task in any text mining pipeline. Much existing work shares one limitation, in that their model training prerequisite and/or runtime computation are too expensive to be applied to all ambiguous entities in real-time. We propose an automatic, light-weight method that processes MEDLINE abstracts at large-scale and with high-quality output. Our method exploits MeSH terms and knowledge in UMLS to first identify unambiguous anchor entities, and then disambiguate remaining entities via heuristics. Experiments showed that our method is 79.6% and 87.7% accurate under strict and relaxed rating schemes, respectively. When compared to MetaMap's disambiguation, our method is one order of magnitude faster with a slight advantage in accuracy.

1 Introduction

Motivation

The ever-growing volume of biomedical literature is published at a phenomenal pace. While the rich information buried in this literature can be extracted via text mining, entity recognition and entity disambiguation – both early tasks in a text mining pipeline – remain challenging. The ideal solution must not only address the quality of the results, but also cope with the sheer volume of textual input. Moreover, the solution should be able to tackle the full spectrum of entities without limiting its scope to narrow specializations such as genes, chemicals, and diseases. For information extraction tasks such as relation mining and knowledge base construction, it is crucial to go beyond merely recognizing entities and strive for precise entities via disambiguation. In this work, we focus on the entity disambiguation task, and propose a solution that attempts to balance quality with high throughput while addressing all entities.

Most existing biomedical entity disambiguation methods that do address all entities cannot be applied in practice to a large corpus for several reasons. The methods based on machine learning (such as Jimeno-Yepes (2016), Chen et al. (2013), Savova et al. (2008), Stevenson et al. (2008)) must identify in advance the exhaustive list of all ambiguous entity names. Where the training is supervised, labeled examples must be obtained, either by expensive manual annotation or by automatic curation (Jimeno-Yepes and Aronson, 2010). Finally, models – in general, one model per ambiguous entity name – must be trained prior to the disambiguation at runtime. All these setup costs render the methods impractical when all ambiguous entity names must be addressed. Alternative methods by Zheng et al. (2015) and Agirre et al. (2010) generate, at runtime, an entire instance of the problem customized per input text. MetaMap (Aronson and Lang, 2010), the de facto standard software tool, disambiguates amongst all entity types but the software is too slow for large-scale usage.

Approach and contributions

We present an automatic and light-weight method that disambiguates all entities in an indexed document by exploiting the indexing as well as domain knowledge. Specifically, the indexed documents are MEDLINE abstracts, which are the bulk of scientific literature in the biomedical domain. As for domain knowledge, the method draws upon UMLS (Unified Medical Language System). We choose MEDLINE abstracts and UMLS as our corpus and knowledge base for this work, respectively, because our method can then leverage the following unique characteristics of these biomedi-

Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pages 72–76,
Berlin, Germany, August 12, 2016.

cal resources:

- MEDLINE abstracts are a large corpus indexed with rich, manually assigned MeSH (Medical Subject Heading) terms; we safely consider all MeSH terms to be accurate. In addition, since abstracts are very compactly written, their content rarely strays away from the biomedical domain. In other words, non-biomedical entities occur only rarely.
- UMLS is the authoritative and comprehensive knowledge base of the biomedical domain covering all aspects of the domain, with a vast collection of entities plus their lexical variations, semantic types, and inter-relationships.
- MeSH terms are themselves a crisp ontology that is already part of UMLS.

Putting these together: All the entities found in a MEDLINE abstract are of a biomedical nature, and all of them can be disambiguated to some canonical entity in UMLS. Therefore, given an abstract, its MeSH terms as ground truth, and all the text mentions in the abstract, the method first identifies unambiguous entities that we shall call *anchors*. The remaining text mentions are then disambiguated using heuristics based on linguistic-semantic patterns and knowledge base assets.

Under the best setting, our method achieves an average of 79.6% and 87.7% accuracy using the strict and relaxed rating schemes, respectively. To the best of our knowledge, this is the first work in the biomedical domain that evaluates all text mentions found in an abstract. In terms of throughput, our method processes 240k abstracts containing 24.5m text mentions in 400 minutes. We also present evaluations against established gold standards via a comparison to MetaMap.

The code is available as an open source project at `http://resources.mpi-inf.mpg.de/d5/bebe/`.

2 Related Work

In the biomedical domain, the terms entity disambiguation and word sense disambiguation are often used interchangeably, since the distinction between entity and sense is not always clear-cut. As mentioned in the Introduction, machine learning-based methods, both supervised and unsupervised, dominate existing works that address all entity types. Domain knowledge is a popular ingredient as well. The most recent work by Jimeno-Yepes (2016) combines word embeddings with long short term memory in a recurrent neural network model. The construction of a custom knowledge graph is the backbone of a collective inference approach by Zheng et al. (2015), where the approach disambiguates multiple entities simultaneously. Chen et al. (2013) applies active learning to support vector machine (SVM). Personalized PageRank is studied by Agirre et al. (2010), relying on and comparing different subsets of UMLS. Four further methods are compared by Jimeno-Yepes and Aronson (2010). In terms of evaluation, two gold standards, NLM WSD (Weeber et al., 2001) and MSH WSD (Jimeno-Yepes et al., 2011), are available.

When it comes to disambiguating only specific or highly specialized entities, a large body of work exists. To name a few representative specializations, there are works that disambiguate between species of genes (Harmston et al., 2012; Wang et al., 2010); chemicals (Batista-Navarro et al., 2015; Leaman et al., 2015); diseases (D'Souza and Ng, 2015); entities in clinical notes (Kang et al., 2012); and coarse entity types (Siu and Weikum, 2015; Jindal and Roth, 2013; Cohen et al., 2011).

3 Methodology

The input to the proposed method is a MEDLINE abstract and its MeSH terms. We use a fast dictionary-based entity recognition tool (Siu et al., 2013) to identify all longest text mentions that match UMLS entity names. (In this work, we use only the license level 0 subset of UMLS, but the proposed method works the same way for larger subsets.) Then the method proceeds in two phases: Phase 1 identifies unambiguous anchors amongst the text mentions. Phase 2 applies heuristics to disambiguate the remaining text mentions.

Unambiguous anchors

In phase 1, the method identifies *anchors* – given a text mention, the method determines if there is one UMLS entity that underlies this text mention unambiguously. A text mention may become an anchor in two ways:

- MeSH term (MESH): Recall that we assume MeSH terms are accurate ground truth. Following a strategy similar to Jimeno-Yepes et al. (2011), this heuristic identifies text mentions that are also MeSH terms for the abstract.
- Only one UMLS match (ONE): Recall that we assume UMLS has complete coverage of

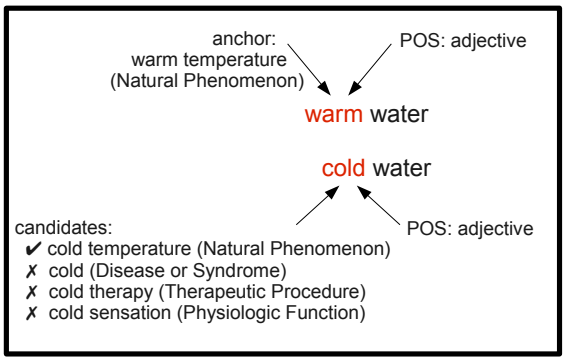

Figure 1: The linguistic-semantic pattern heuristic

all biomedical entities. A text mention that matches only a single UMLS entity is therefore considered unambiguous.

We pin down the anchors so that their underlying entities are considered correctly disambiguated.

Heuristics

In phase 2, the method disambiguates any remaining, non-anchor text mentions. Recall that the entity recognition tool already provides multiple matching UMLS entities to such a text mention. Taking these UMLS entities as candidates, select one candidate using one or more heuristics:

- Singular/plural (SP): Since abstracts are very short documents, we assume that, within one abstract, text mentions sharing the same surface string also share the same entity. Therefore, singular (e.g. *diet*) and plural (*diets*) forms of the same word should refer to the same entity. In UMLS, when the plural form is a unique entity (C0012155), that same entity is extended to the singular form, and vice versa.
- Linguistic-semantic pattern (PAT): Figure 1 depicts this heuristic via the example of two bigrams, *warm water* and *cold water*. When one word (*water*) appears in both bigrams in the same position, and when the other words (*warm* and *cold*) have the same part-of-speech, *warm* and *cold* ought to share the same linguistic function and some analogous meaning. Since *warm* is an anchor, take its UMLS semantic type (Natural Phenomenon), and pick for *cold* a candidate with the same type (cold temperature the Natural Phenomenon).
- Co-occurring semantic types (CO): The intuition behind this heuristic is that objects of the same semantic type often co-occur in the same abstract. For instance, an abstract mentioning different fish species naturally also mentions the word *fish*. However, the candidates for *fish* belong to different UMLS semantic types (Fish, Gene or Genome, Organic Chemical (for fish extract), and Molecular Biology Research Technique (for Fluorescence in situ Hybridization)). When the entities in the abstract exhibit a predominant semantic type, pick the candidate with the same type.
- Ranked preferences of dictionary sources (RANK): UMLS comes with a pre-defined preference list of dictionary sources; more specifically, the source attributes have numerical ranks in the MRRANK table. When a text mention matches multiple candidates, each candidate's dictionary source leads to a corresponding rank number. This heuristic picks the candidate with the best rank. Under this heuristic, for instance, *HIV* the virus is preferred over *HIV* the vaccine.
- Prior probability (PRIOR): Thanks to the heterogeneous nature of UMLS, the listing of entity names contains much redundancy. Specifically, a single entity name is listed separately for each dictionary's contribution. A more popular meaning of the word (e.g. *cat* the animal) appears in more rows of the MRCONSO table than a less popular meaning (*CAT* the scan procedure). The prior probability distribution of candidates is thus estimated based on counts of entity name occurrences. Our prior work (Siu and Weikum, 2015) shows that estimated prior probabilities contribute to enriching disambiguation contexts 72% of the time. Here, the heuristic picks the candidate with the highest prior probability.

4 Results and Discussion

Ablation study of heuristics

We used disjoint sets of MEDLINE abstracts published in 2014 as the development and test datasets. The test dataset, in particular, consists of 20 randomly selected abstracts; in total, 2,549 text mentions were recognized. Two annotators evaluated all the recognized text mentions, including the anchors, rating the candidates as "completely correct", "partially correct", or "completely wrong". The inner-annotator agreement, calculated as Cohen's kappa, was 0.64, which in-

Heuristic(s)	Anchors Strict	Anchors Relaxed	Non-anchors Strict	Non-anchors Relaxed	All text mentions Strict	All text mentions Relaxed
MESH	16.0%	16.9%	not applicable		8.9%	9.4%
ONE	83.3%	85.0%			46.5%	47.5%
MESH + ONE	90.3%	93.0%			50.4%	51.9%
MESH + ONE + CO	remains at 90.3%	93.0%	47.2%	72.3%	71.3%	83.9%
MESH + ONE + PAT			7.8%	9.9%	53.9%	56.3%
MESH + ONE + PRIOR			53.9%	68.9%	74.2%	82.4%
MESH + ONE + RANK			63.2%	77.9%	78.5%	86.3%
MESH + ONE + SP			18.6%	22.1%	58.6%	61.7%
Successive filtering	remains at 90.3%	93.0%	66.2%	79.0%	**79.6%**	86.8%
Majority voting			64.3%	81.1%	78.8%	**87.7%**

Table 1: Contribution of different heuristics to accuracy

dicates mostly substantial agreement. The presence of fine shades of the same underlying entity in UMLS prompted the "partially correct" annotation choice. For instance, *children* exists as two separate entities with the semantic types Age Group and Family Group, and the exact distinction is difficult even for human judges. We therefore present results in two rating schemes: Under the strict rating scheme, only "completely correct" annotations count as correct; under the relaxed scheme, both "completely correct" and "partially correct" annotations count as correct.

Table 1 shows the accuracy and the contribution of each heuristic. We experimented with two types of ensembles, namely majority voting and applying heuristics as successive filters similar to D'Souza and Ng (2015). Under the relaxed rating scheme, majority voting consistently performed better. The best ensemble used, as expected, all heuristics to reach 87.7% accuracy. Under the strict rating scheme, on the other hand, successive heuristic filters consistently performed better. The best ensemble scored 79.6% accuracy using the following order of heuristics: MESH, ONE, SP, RANK, PRIOR, PAT, CO. On average, 56% of all text mentions in an abstract were anchors.

Comparison with MetaMap and other datasets

We compared the best setting of our method with MetaMap (version 2016 with disambiguation) using the aforementioned custom test dataset as well as 3 other datasets: NLM WSD (Weeber et al., 2001), EBI disease corpus (Jimeno et al., 2008), and a subset of the CRAFT corpus (Bada et al., 2012) that provides UMLS entity IDs in abstracts. Table 2 shows the accuracy for both systems. (The MSH WSD dataset (Jimeno-Yepes et al., 2011) was not used here because it was essentially constructed with the MESH heuristic; using the dataset would not offer further insight.)

	Custom strict	Custom relaxed	NLM WSD	EBI disease	CRAFT subset
Our method	79.6%	87.7%	39.9%	87.3%	38.8%
MetaMap	68.1%	76.1%	33.7%	78.4%	33.0%

Table 2: Comparison of accuracy between our method and MetaMap

In terms of accuracy, both systems showed analogous trends for each dataset, though our proposed method outperformed MetaMap by 5% to 11%. Both systems performed poorly over the NLM WSD and CRAFT datasets due to their wide variety of highly ambiguous entity names. The disambiguation module in MetaMap is known to be a weaker module in the system (Aronson and Lang, 2010), while our method's heuristics are too simplistic for sophisticated cases. The same rationale explains why accuracy in EBI disease corpus was high, because disease names are much less ambiguous in general. In terms of speed, our system and MetaMap processed 600 and 11 abstracts per minute, respectively, on the same linux machine with 8 Intel Xeon 2.4GHz CPUs and 48GB RAM.

5 Conclusions and Future Work

We present a large-scale, high-quality, and automatic method that disambiguates entities in MEDLINE abstracts by exploiting MeSH terms as well as applying heuristics based on linguistic cues and knowledge assets in UMLS. Not only is the proposed method one order of magnitude faster than MetaMap, the overall accuracy is also slightly superior to that of MetaMap. Therefore we further propose our method as a viable alternative for real-time processing. We plan to harness the outputs of this work for future investigation on biomedical entity disambiguation.

References

Eneko Agirre, Aitor Soroa, and Mark Stevenson. 2010. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

Alan R. Aronson and François-Michel Lang. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A. Baumgartner, K. Bretonnel Cohen, Karin Verspoor, Judith A. Blake, and Lawrence E. Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):1–20.

Riza Theresa Batista-Navarro, Rafal Rak, and Sophia Ananiadou. 2015. Optimising chemical named entity recognition with pre-processing analytics, knowledge-rich features and heuristics. *Journal of Cheminformatics*, 7(S-1):S6.

Yukun Chen, Hongxin Cao, Qiaozhu Mei, Kai Zheng, and Hua Xu. 2013. Applying active learning to supervised word sense disambiguation in MEDLINE. *Journal of the American Medical Informatics Association*, 20(5):1001–1006.

Raphael Cohen, Avitan Gefen, Michael Elhadad, and Ohad S. Birk. 2011. CSI-OMIM – Clinical synopsis search in OMIM. *BMC Bioinformatics*, 12:65.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL) (Volume 2: Short Papers)*, pages 297–302.

Nathan Harmston, Wendy Filsell, and Michael Stumpf. 2012. Which species is it? Species-driven gene name disambiguation using random walks over a mixture of adjacency matrices. *Bioinformatics*, 28(2):254–260.

Antonio Jimeno-Yepes, Ernesto Jimenez-Ruiz, Vivian Lee, Sylvain Gaudan, Rafael Berlanga, and Dietrich Rebholz-Schuhmann. 2008. Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics*, 9(3):1–10.

Antonio Jimeno-Yepes and Alan R. Aronson. 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics*, 11(1):1–12.

Antonio Jimeno-Yepes, Bridget McInnes, and Alan R. Aronson. 2011. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics*, 12:223.

Antonio Jimeno-Yepes. 2016. Higher order features and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. *arXiv preprint arXiv:1604.02506*.

Prateek Jindal and Dan Roth. 2013. Using soft constraints in joint inference for clinical concept recognition. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1808–1814.

Ning Kang, Zubair Afzal, Bharat Singh, Erik M. van Mulligen, and Jan A. Kors. 2012. Using an ensemble system to improve concept extraction from clinical records. *Journal of Biomedical Informatics*, 45(3):423 – 428.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong Lu. 2015. tmChem: a high performance approach for chemical named entity recognition and normalization. *Journal of cheminformatics*, 7(supplement 1).

Guergana K. Savova, Anni R. Coden, Igor L. Sominsky, Rie Johnson, Philip V. Ogren, Piet C. de Groen, and Christopher G. Chute. 2008. Word sense disambiguation across two domains: biomedical literature and clinical notes. *Journal of Biomedical Informatics*, 41(6):1088–1100.

Amy Siu and Gerhard Weikum. 2015. Semantic type classification of common words in biomedical noun phrases. In *Proceedings of BioNLP 2015*, pages 98–103.

Amy Siu, Dat Ba Nguyen, and Gerhard Weikum. 2013. Fast entity recognition in biomedical text. In *Workshop on Data Mining for Healthcare (DMH) at Knowledge Discovery and Data Mining (KDD) 2013*.

Mark Stevenson, Yikun Guo, Robert Gaizauskas, and David Martinez. 2008. Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics*, 9(Suppl 11):S7.

Xinglong Wang, Jun'ichi Tsujii, and Sophia Ananiadou. 2010. Disambiguating the species of biomedical named entities using natural language parsers. *Bioinformatics*, 26(5):661–667.

Marc Weeber, James G. Mork, and Alan R. Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMIA Symposium*, pages 746–750.

Jin G. Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2015. Entity linking for biomedical literature. *BMC Medical Informatics and Decision Making*, 15(1):1–9.