Cross-Evaluation of Entity Linking and Disambiguation Systems for Clinical Text Annotation

Camilo Thorne
Universität Mannheim
camilo@informatik.uni-mannheim.de

Stefano Faralli
Universität Mannheim
stefano@informatik.uni-mannheim.de

Heiner Stuckenschmidt
Universität Mannheim
heiner@informatik.uni-mannheim.de

ABSTRACT

In this paper we study whether state-of-the-art techniques for multi-domain and multilingual entity linking can be ported to the clinical domain. To do so, we compare two known entity linking systems, BabelFly and TagMe, that leverage on Wikipedia and DBpedia, with the standard clinical semantic annotation and disambiguation system, MetaMap, over the SemRep clinical word sense disambiguation gold standard. We show that BabelFly and especially TagMe, while achieving decent precision on clinical annotation, outmatch MetaMap's $F1$-score.

CCS Concepts

•Information systems → Information extraction; •Computing methodologies → Lexical semantics; Ontology engineering;

Keywords

Entity linking; semantic annotation; DBpedia; clinical NLP.

1. INTRODUCTION

Nowadays, textual information increasingly pervades healthcare. From electronic health records, to clinical papers, online dictionaries or clinical guidelines, an ever larger share of knowledge underpinning clinical decision-making is available as raw, unstructured text [5]. This has motivated the application and adaptation of natural language processing (NLP) techniques to index and search, but also to extract knowledge from clinical text.

An important task in information extraction is semantic annotation in the form of *entity extraction*[1]. To solve this task we need to perform two steps: **(1)** identify *entities* (nouns, noun phrases) within an text; **(2)** identify or resolve the meaning of such (generic) entities within such text by linking them to a sense repository. Consider the following example:

Low dose **pramipexole** is neuroprotective in the MPTP mouse model of **Parkinson's disease**. (*)

In (*) three generic entities have been identified, two clinical (bold) and one related to world knowledge. By "Pramipexole" is meant a kind of drug that activates dopamine receptors; by "model" is meant a theoretical framework; and by "Parkinson's disease" the known neurodegenarative disease.

To solve **(1)** and **(2)** we need to solve an *entity linking* (EL) and a *word-sense disambiguation* (WSD) task: to identify the senses $sen(e)$ of a (generic) entity e within a sense inventory and then determine, based on its sentential context, which is the one effectively used. While there are systems and sense inventories that can extract and disambiguate either clinical or world-knowledge entities, the question arises as to whether there are systems capable of *annotating both*.

For the clinical domain, the standard resource is the US national Library of Medicine (NLM) Unified Medical Language System (UMLS) Metathesaurus[2]. For world (encyclopedic) knowledge known resources are, e.g., WordNet [3]. However, more recently, BabelNet [13] has arisen as a multilingual thesaurus that incorporates lexical knowledge extracted from Wikipedia. Moreover, linked open data, in the form of open domain knowledge bases or ontologies such as DBpedia [2], also based on encyclopedic knowledge extracted from Wikipedia, have also been proposed as sense inventories [8] for EL and WSD. So has been Wikipedia itself [7]. It is known that Wikipedia covers a significant portion of clinical senses [14]. Thus, Wikipedia-driven annotators appear like good candidates for annotating both clinical and non-clinical entities, but their performance in the clinical domain remains relatively unknown.

The contributions of this paper are three: **(1)** We provide a comparative evaluation of clinical annotation resources (MetaMap) and two state-of-the-art Wikipedia-driven linking systems, BabelFly and TagMe, over a known clinical gold WSD corpus, the SemRep corpus. **(2)** We exploit open linked data to align all senses (gold senses and annotator senses) to DBpedia. **(3)** We study whether the annotations returned by each annotator are semantically similar among themselves and with respect to the gold corpus.

2. SENSE REPOSITORIES

UMLS. The UMLS Metathesaurus is the main thesaurus used for the clinical domain. Developed since the early 90's by the NLM[3], it is built around the notion of a *clinical concept*, to which a so-called *concept unique identifier* (CUI) is associated. Concepts are organized into semantic types, and related by a certain number semantic relations, giving rise to a so-called semantic network.

WordNet. WordNet, developed in the 90s at Princeton, is a well-known sense inventory for standard English. It is organized around

[1] We assume following Hobbs in [6] a broader notion of *entity* that encompasses not only names, but any *noun* or *noun phrase* in a sentence.

SEMANTiCS 2016, September 12-15, 2016, Leipzig,Germany

DOI: http://dx.doi.org/10.1145/2993318.2993345

[2] *https://www.nlm.nih.gov/research/umls/*

[3] *https://uts.nlm.nih.gov*

	sense	sense ID	DBpedia URI
Clinical (Gold)	pramipexol	C0074710	*http://dbpedia.org/resource/Pramipexole*
	Parkinson disease	C0030567	*http://dbpedia.org/resource/Parkinson_disease*
MetaMap	pramipexol	C0074710	*http://dbpedia.org/resource/Pramipexole*
	Parkinson disease	C0030567	*http://dbpedia.org/resource/Parkinson_disease*
BabelFly	ATC_code_N04BC05	bn:03124207n	*http://dbpedia.org/resource/Pramipexole*
TagMe	pramipexole	*https://en.wikipedia.org/wiki/Pramipexole*	*http://dbpedia.org/resource/Pramipexole*
	Parkinson's disease	*https://en.wikipedia.org/wiki/Parkinson's_disease*	*http://dbpedia.org/resource/Parkinson's_disease*

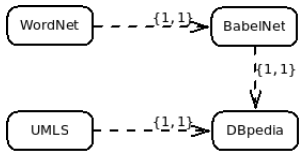


Figure 1: Left: Semantic *unresolved* annotations for example (*). On the first two rows, gold standard clinical senses, together with their UMLS CUI and the corresponding DBpedia URIs. Next, corresponding MetaMap, BabelFly and TagMe annotations. MetaMap senses are UMLS CUIs. BabelFly senses, BabelNet synset IDs. The WordNet baseline returned no links for either (generic) entity. Right: Outline of the mappings among the resources. While only partial they are (1) one-to-one and (2) pivot around DBpedia.

the notion of word *sense*, viz., a context that uniquely characterizes a particular use of a given word or term. Senses are organized into *synsets*, a class of synonymous senses, among which hold a number of so-called *lexical relations*, such as, synonymy, antonymy, hypernonymy, homonymy, etc. Senses and synsets are defined via *glosses*. Currently, WordNet[4] comprises 117,000 synsets.

Wikipedia. Wikipedia[5] is an online multilingual, cross-domain encyclopedia covering around 292 languages (282 active) and currently 5,145,674 articles for English alone. It is loosely structured around a taxonomy of topics. Wikipedia has been used a resource for (generic) entity recognition, linking and extraction as a large number of its articles define or describe entities.

DBpedia. DBpedia [2] on the other hand, emerged in the 2000s partly as an attempt to add further structure to Wikipedia, via semantic technologies (the Resource Description Framework or RDF data model) and open information extraction techniques. DBpedia is built around *resources*, encoding Wikipedia entries, each identified by a *unique resource identifier* (URI). Resources are organized into a knowledge base, viz., into triples describing resource typing and binary relations among resources, but also intensional knowledge over such types and relations. DBpedia covers 111 languages, and 65,143,840 extensional triples for English alone[6].

BabelNet. BabelNet [13] was built by automatically extending WordNet with lexical knowledge gathered from Wikipedia, from several multilingual extensions of WordNet and WordNets developed for languages other than English. As such, it extends WordNet via the notion of a multilingual *babelsynset*, which, in addition to multilinguality, contains links to DBpedia resources in the form of DBpedia resource URIs. Currently, BabelNet[7] covers around 14 million babelsynsets, and 271 languages.

3. ANNOTATION EVALUATION

Sense Linking. Two semantic resources T and T', are said to *align* when there exists a mapping $f: T \rightarrow T'$ that assigns to each sense $s \in T$ a unique sense $f(s) \in T'$. Such mappings, when they exist, tend to be partial at best. They are however critical for evaluating results among annotators that rely on different semantic resources, in particular when they give rise to a "pivot" resource into which they can be all mapped. It is known that world-knowledge resources can be mapped in such manner to DBpedia. Interestingly, this is also possible, albeit in a more restricted form, for the UMLS.

BabelNet strictly extends WordNet [13], giving rise to an injective, total mapping of WordNet synsets to BabelNet babelsynsets. Babelsynsets in their turn can be partially mapped into DBpedia, and assigned a DBpedia resource URI. Finally, DBpedia contains a resource URI for each Wikipedia entry [2].

Regarding UMLS CUIs, the mapping is less obvious, but still possible. In [11] it has been shown that UMLS CUIs can be partially aligned with DBpedia URIs via Medical Subject Headings (MeSH) terms. MeSH is a controlled vocabulary of subject headings and related glosses (short descriptions or definitions). Most UMLS CUIs can be associated to a unique MeSH term, which in its turn, can be linked or mapped to a DBpedia URI. This mapping has given rise to a linked open data knowledge base, Liked Life Data[8] that can be used to partially align the UMLS with WordNet, BabelNet and DBpedia. Figure 1, right, shows the relationships among the different resources.

3.1 Annotation and Linking Systems

WordNet (baseline). As customary for systems that rely on WSD, we implemented as baseline a WordNet-based annotator using the known Lesk WSD algorithm [10]. The Lesk algorithm maximizes the bag-of-words similarity measure between entity e's sentence S (represented as a bag of words) and the glosses of its candidate senses (WordNet synsets) $s \in sen(e)$.

MetaMap. MetaMap [1] is a state-of-the-art clinical text semantic annotator developed also by the NLM. MetaMap recognizes and disambiguates clinical terms in texts, associating to them its most likely UMLS CUI. MetaMap does not really exploit sentence context, but relies on a series of similarity metrics based to a large degree on morphological analysis to match entities to UMLS concepts. MetaMap is available as a (RESTFul) webservice[9].

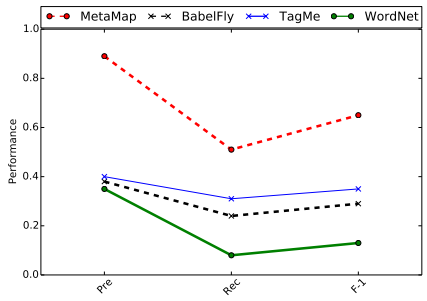
BabelFly. BabelFly [12] is a state-of-the-art semantic annotator and entity linker that assigns to each (generic) entity e a BabelNet babelsynset s. Similarly to Lesk, it also relies on sentence context, but instead of bag-of-words similarity it exploits the graph structure of thesauri (induced by lexical relations of synonymy, hypernonymy meronymy, etc.) to maximize a measure of *clique* similarity. As a side effect, it also returns the DBpedia URI associated to s as well as its lemmas in all the languages covered by the babelsynset. BabelFly is available as a (RESTFul) webservice[10].

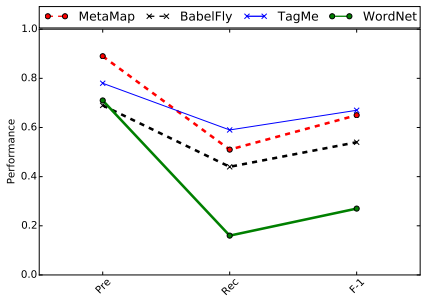
Tagme. TagMe [4] is an annotator that links words to Wikipedia pages (i.e., to Wikipedia page URLs). TagMe works, broadly speaking, by jointly maximizing the *coherence* of all the (generic) entities $e_1, \ldots, e_k$ spotted in a sentence S. By coherence is understood a measure of topical relatedness among such terms and Wikipedia entries. TagMe covers currently two languages, English an Italian, and is also available as a (RESTFul) web service[11].

3.2 Evaluation

The SemRep corpus. We ran our experiments over the SemRep corpus [9], a small annotated clinical corpus consisting of 428 clin-

[4] *https://wordnet.princeton.edu* [5] *http://wikipedia.com*
[6] *http://dbpedia.org* [7] *http://babelnet.org*
[8] *http://linkedlifedata.com* [9] *https://metamap.nlm.nih.gov*
[10] *http://babelfly.org* [11] *http://tagme.di.unipi.it*





Exp.	Test	Null hyp.	p-value
Unres.	Kruskal-Wallis	identical	0.051
Res.	Kruskal-Wallis	identical	0.056

Test	Annotator	Null hyp.	p-value
t-test	BabelFly	identical	0.015
t-test	TagMe	identical	0.008
t-test	WordNet	identical	0.151

Figure 2: Left: Annotation performance with *unresolved* DBpedia URIs. Middle: Annotation performance with *resolved* DBpedia URIs. Right, top: Statistical significance analysis of results; comparison of *Pre*, *Rec* and *F1* among MetaMap, BabelFly, TagMe and WordNet baseline. Right, bottom: Statistical significance analysis of URI resolution; BabelFly and TagMe show a strongly significant improvement.

ical excerpts (MedLine/PubMed) and 13,948 word tokens, roughly representative of the whole clinical domain. For each sentence, two noun phrases were annotated with their corresponding UMLS CUI by clinicians and domain experts, giving rise to a total of 856 UMLS-annotated clinical entities. Of these, 606 can be associated to a corresponding DBpedia URI. See Figure 3, left.

Annotation procedure. We ran the 4 annotators separately over the SemRep corpus. For each of the annotators (WordNet baseline, MetaMap, BabelFly, TagMe) we collected two kinds of information: **(1)** the corresponding entity sense annotations for the Gold corpus entities, together with their associated DBpedia URIs, whenever defined, viz., two per sentence; **(2)** for each sentence in the corpus, the senses of all the sentence's entities, their DBpedia URIs and their glosses[12]. See Figure 1, left, to get a glimpse as to how such annotations look like.

URI resolution. Different annotators may return senses that are only superficially different, and constitute only notational or lexical variants of the same synset or concept. In Figure 1, left, MetaMap returns *http://dbpedia.org/resource/Parkinson_Disease*, and TagMe returns *http://dbpedia.org/resource/Parkinson's_Disease*, two superficially different albeit identical URIs *modulo identity*. To understand how much impact this phenomenon has on annotator performance we additionally resolved such identities. Specifically, we tested for `owl:sameAs` and `dbpedia:redirects` relationships among the corresponding DBpedia URIs. On average, we discovered that one out of three of the URIs associated either to the corpus' CUIs or to the senses returned by the annotators is a variant (*modulo* `owl:sameAs` and `dbpedia:redirects`) of a previously introduced DBpedia URI. See Figure 3, left.

Annotation performance. We evaluated MetaMap, BabelFly, TagMe and our Lesk-driven WordNet WSD baseline over the 606 DBpedia-annotated entities of the SemRep gold standard, and measured their *precision* (*Pre*), *recall* (*Rec*) and *F1-measure* (*F1*)[13], w.r.t. the unresolved and the resolved DBpedia URIs.

The results of the evaluation are summarized by Figure 1. As expected, MetaMap achieved the highest precision, although at the expense of relatively low recall –over the clinical domain itself–: $Pre = 0.89$, $Rec = 51$, $F1 = 65$. Regarding the other annotators, again, as expected, our WordNet baseline performed uniformly worst, but TagMe surprisingly achieved better results than BabelFly, and even slightly surpassed the *F1* score of MetaMap (0.67) when URIs where resolved (while achieving 0.78 precision)[14]. Such differences were close to significant. Moreover, as the reader can see, while resolving URIs had little to no impact on MetaMap, it did improve significantly the results of all the other annotators[15].

Gloss similarity performance. Given a clinical (generic) entity e annotated with sense s in the SemRep corpus, we measured the similarity of the corresponding sense s' returned by MetaMap, BabelFly, TagMe and WordNet by measuring how many times they were **(1)** WordNet related and **(2)** distributionally related ; we considered moreover two relatedness thresholds, a "strict" threshold > 0.2 and a "loose" threshold > 0. This gave rise to the following four similarity metrics. Respectively, WordNet similarity, WordNet loose similarity, distributional similarity and distributional loose similarity:

$$syn(s,s') = \frac{\sum\{(w,w') \in g(s) \times g(s') \mid wn_{>0.2}(w,w')\}}{|g(s)| + |g(s')|}$$

$$syn^{+}(s,s') = \frac{\sum\{(w,w') \in g(s) \times g(s') \mid wn_{>0}(w,w')\}}{|g(s)| + |g(s')|}$$

$$dsyn(s,s') = \frac{\sum\{(w,w') \in g(s) \times g(s') \mid dn_{>0.2}(w,w')\}}{|g(s)| + |g(s')|}$$

$$dsyn^{+}(s,s') = \frac{\sum\{(w,w') \in g(s) \times g(s') \mid dn_{>0}(w,w')\}}{|g(s)| + |g(s')|}$$

where $g(s)$ denotes the bag-of-words derived from s' gloss. For distributional similarity, we used the standard Wikipedia-trained word2vec word space models. For WordNet and BabelFly we retrieved the corresponding WordNet and BabelNet glosses of the returned sense. For TagMe, we considered as "gloss" (sense definition) the DBpedia resource URI's abstract. Finally, for MetaMap and the SemRep corpus itself, we considered as "glosses" the MeSH (or Medline) abstract LinkedLifeData assigns CUIs.

As the reader can see in Figure 3 (middle and right), this analysis gave rise to only slight differences among the annotators. Such differences were not statistically significant.

Discussion. The annotation results show that, while MetaMap (and arguably, similar UMLS-annotators) does indeed beat general purpose entity linking systems on precision these perform reasonably well, providing in addition broad coverage, in particular when URIs are resolved. TagMe even slighly surpasses in F1-score MetaMap

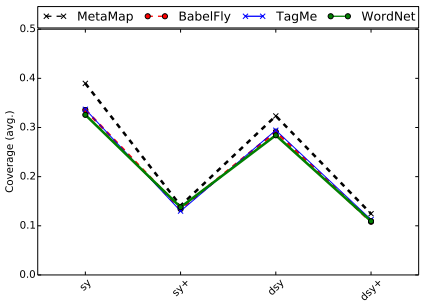
[12] BabelFly and TagMe were called with their default parameters.

[13] Where: $Pre = \frac{\#\text{corrent senses}}{\#\text{returned senses}}$, $Rec = \frac{\#\text{corrent senses}}{\#\text{corpus senses}}$, $F1 = \frac{2 \cdot Pre \cdot Rec}{Pre + Rec}$.

[14] BabelFly achieved 0.69 precision and a F1-score of 0.54 by comparison.

[15] We report only those tests that yielded the lowest p-values.

# of CUIs in corpus (total)	=	856
# of corpus DBpedia URIs	=	606
# of resolved corpus URIs	=	404
# of MetaMap DBpedia URIs	=	343
# of resolved MetaMap URIs	=	242
# of BabelFly DBpedia URIs	=	432
# of resolved BabelFly URIs	=	269
# of TagMe DBpedia URIs	=	469
# of resolved TagMe URIs	=	320
# of WordNet DBpedia URIs	=	182
# of resolved WordNet URIs	=	97



Annotations	Avg. len. (per sent.)
Corpus sense glosses	66.41 words
BabelFly sense glosses	199.43 words
TagMe sense glosses	325.51 words
MetaMap sense glosses	191.76 words
WordNet sense glosses	50.50 words

Test	Null hyp.	p-value
Kruskal-Wallis	identical	0.897

Figure 3: Left: Annotation statistics. On the first row, the total number of gold standard CUIs in SemRep. Next, the number of CUIs with an associated DBpedia URI (resp. a *resolved* URI). Finally, the number of annotations (resp. resolved) annotations returned by each annotator over the 606 DBpedia-linked entities of SemRep. Middle: Gloss similarity across annotators. Right, top: Average length of glosses (in content words), per sentence (the glosses were computed for all the noun phrases). Right, bottom: A Kruskal-Wallis test showed no significant change between annotators.

(and all other annotators). We think that this is due to **(1)** their more advanced WSD algorithms, and, crucially, **(2)** the fact that they leverage encyclopedic knowledge that largely covers the clinical domain.

The results of the similarity evaluation show, on the other hand, that **(3)** sense glosses appear to be clinical domain related, but **(4)** are not sufficiently informative to distinguish the annotators. This may be due to the verbosity of the glosses, which (at a average, rate of 166.22 content words per sentence, all entities and annotators taken together), while broadly speaking remaining in the clinical domain, contain information not directly relevant for a term's sense definition.

4. CONCLUSIONS

We have carried out a cross-evaluation of domain specific (clinical) and cross- and multiple-domain semantic annotation systems over the SemRep clinical word sense disambiguation gold standard, to answer the question whether generic state-of-the-art systems can match in performance domain specific tools, while being able to semantically annotate *both* domain-specific and world-knowledge entities. In particular, we have compared two known entity linking systems, BabelFly and TagMe, that leverage on Wikipedia and DBpedia, plus a simple WordNet baseline, with the standard clinical semantic annotation and disambiguation system, MetaMap. Our results show that BabelFly and TagMe perform reasonably well over the clinical domain (their precision lying close to that of MetaMap), while achieving a much wider coverage (higher recall) than MetaMap. This is due to their leveraging on the large body of encyclopedic knowledge from Wikipedia, as the poor results of the baseline (restricted to WordNet) show.

5. REFERENCES

[1] A. R. Aronson and F.-M. Lang. And overview of MetaMap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. DBpedia - A crystallization point for the web of data. *Journal of Web Semantics*, 7(3):154–165, 2009.

[3] e. Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[4] P. Ferragina and U. Scaiella. TAGME: on-the-fly annotation of short text fragments (by wikipedia entities). In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management (CIKM 2010)*, 2010.

[5] G. Goth. Analyzing medical data. *Communications of the ACM*, 55(6):13–15, 2012.

[6] J. R. Hobbs. The generic information extraction system. In *Proceedings of the Fifth Message Understanding Conference (MUC-5)*, 1993.

[7] E. Hovy, R. Navigli, and S. P. Ponzetto. Collaboratively built semi-structured content and artificial intelligence: The story so far. *Artificial Intelligence*, 194:2–27, 2013.

[8] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene. An eigenvalue-based measure for word-sense disambiguation. In *Proceedings of the 25th International Florida Artificial Intelligence (FLAIRS 2012)*, 2012.

[9] H. Kilicoglu, G. Rosenblat, M. Fiszman, and T. C. Rindfleisch. Constructing a semantic predication gold standard from the biomedical literature. *BMC Bioinformatics*, 12(486), 2011.

[10] M. Lesk. Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation (SIGDOC 1986)*, 1986.

[11] V. Momtchev, D. Peychev, T. Primov, and G. Georgiev. Expanding the pathway and interaction knowledge in linked life data. *Proceedings of 2009 International Semantic Web Challenge*, 2009.

[12] A. Moro, A. Raganato, and R. Navigli. Entity linking meets word sense disambiguation: a unified approach. *Transactions of the Association for Computational Linguistics*, 2:231–244, 2014.

[13] R. Navigli and S. P. Ponzetto. BabelNet: Building a very large multilingual semantic network. In *In Proceedings of the 48th Annuanl Meeting og the Association for Computational Linguistics (ACL 2010)*, 2010.

[14] A. Sajadi, E. E. Milios, V. Keselj, and J. C. M. Janssen. Domain-specific semantic relatedness from Wikipedia structure: A case study in biomedical text. In *Proceedings of the 16th International Conference in Computational Linguistics and Intelligent Text Processing (CICLing 2015)*, 2015.