Preliminary results for Biomedical Word Sense Disambiguation based on Semantic Clustering

Tamara Martín-Wanton
Universitat Jaume I
Castellón, Spain
tmartin@guest.uji.es

Rafael Berlanga-Llavori
Universitat Jaume I
Castellón, Spain
berlanga@lsi.uji.es

Antonio Jimeno-Yepes
National Library of Medicine
Bethesda, MD 20894, USA
antonio.jimeno@gmail.com

***Abstract*—Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, attempting to select the proper sense of ambiguous words.**

Due to the scarcity of training data, knowledge-based and knowledge-lean methods receive attention as disambiguation methods. Knowledge-based methods compare the context of the ambiguous word to the information available in the terminological resource, but their main purpose is not only word sense disambiguation. Knowledge-lean unsupervised methods rely on terms distribution instead of a resource enumerating the possible senses but might be inappropriate when there is a requirement to commit to a terminological resource.

In this work, we rely on a Knowledge Resource (KR) which provides both an inventory of concepts and their lexical information. Our aim is to design scalable unsupervised WSD methods for the semantic annotation of large biomedical corpora. More specifically, we present a clustering-based method that takes profit from the KR information encoded in form of kernels. Prelimanary results are compared to state-of-the-art methods for unsupervised WSD.

***Keywords*-word sense disambiguation; clustering;**

I. Introduction

Word sense disambiguation (WSD) is an intermediate task within information retrieval and information extraction, attempting to select the proper sense of ambiguous words. For instance, the word *cold* could either refer to *low temperature* or *viral infection*.

Several methods to perform WSD from supervised to knowledge-based approaches rely on a resource enumerating words and their possible senses. Supervised methods usually achieve the best performance in disambiguation but require training data for each disambiguation example [1]. Knowledge-based methods compare the context of the ambiguous word to the information available in the resource, so do not need training data but achieve lower performance.

Knowledge-lean unsupervised methods rely on term distribution instead of a resource enumerating the possible senses [2]. Usually, these methods first perform a sense discrimination step and then a sense labeling step. LDA (Latent Dirichlet Allocation) [3] has been used to provide better results compared to existing approaches. But identified candidate senses might not correlate with all the senses in reference terminological resources. Their main interest is portability of methods to several domains but might not be fully adequate when the disambiguation method requires compliance with senses enumerated in a terminological resource.

Our motivation is to define scalable approaches for disambiguating annotations in large biomedical corpora with acceptable effectiveness. We aim at disambiguate all ambiguous annotations found in text by looking for their affinities. This kind of WSD is required in summarization and information extraction based on a knowledge resource such as the Unified Medical Language System (UMLS).

II. Related work

Scarcity of training data due to its cost makes unsupervised methods more appealing compared to supervised ones. Many knowledge-based algorithms can be seen as a relaxation of Lesk's algorithm [4], which is very expensive since the sense combination might be exponentially large even for a single sentence. Vasilescu et al. [5] have shown that similar or even better performance might be obtained disambiguating each ambiguous word separately. Most of the knowledge-based methods can be broadly divided into two categories, namely, similarity- and graph-based ones [6].

The first category compares each sense of a target word with its surrounding context words. The sense that has the highest similarity is assumed to be the right one. In these approaches, correct senses are determined for each word individually without considering the senses previously assigned. With this aim, different similarity measures have been used such as information content [7], conceptual density [8] and extended gloss overlaps [9]. Budanitsky and Hirts [10] evaluated the performance of a number of measures of semantic relatedness that have been proposed for use in applications in natural language processing and information retrieval.

In graph-based methods [11], [12], [6], a graph whose nodes are word senses and edges represent meaningful relations or dependencies between them, is built from lexical resources. This graph structure is assessed to determine the importance of each node and the correct sense corresponds to the most important node for each word. Unlike the similarity-based approaches, here word senses are globally determined by capturing their relationships. Experimental studies [11], [13] show that graph-based methods outperform similarity-based ones. As in Mihalceas method, in our clustering-based approach we build a weighted graph whose nodes are word senses and edges are labeled with the affinity between them, but instead of determining the importance of a sense by using centrality algorithms, we iteratively perform a clustering method to discover the relationships existing among senses to identify the right ones.

Related methods further use the inner structure of the terminological resource to approximate the sense bias [14], while other approaches additionally use available corpora to recover context examples of the ambiguous word to train supervised learning approaches [15].

We propose to exploit the information provided by the KR in order to find out concept affinities that help us to decide about the most adequate concept. Concepts affinities are used for clustering concepts and then clusters to the context. Thus, our hypothesis is that the correct sense of a word is determined by the correct sense of its neighbours, which must be placed in the same affinity-like cluster similar to the context.

III. UMLS

The NLM's UMLS [16] provides a large resource of knowledge and tools to create, process, retrieve, integrate and/or aggregate biomedical and health data. The UMLS has three main components:

- Metathesaurus, a compendium of biomedical and health content terminological resources under a common representation which contains lexical items for each one of the concepts, relations among them and possibly one or more definitions depending on the concept. In the 2009AB version, it contains over a million concepts.
- Semantic network, which provides a categorization of Metathesaurus concepts into semantic types. In addition, it includes relations among semantic types.
- SPECIALIST lexicon, containing lexical information required for natural language processing which covers commonly occurring English words and biomedical vocabulary.

Concepts are assigned a unique identifier (CUI) which has linked to it a set of terms which denote alternative ways to represent the concept, for instance, in text. Concepts are assigned one or more semantic types.

IV. Kernels

Kernel methods [17], [18] are an attractive alternative to feature-based methods. Kernel methods retain the original representation of objects and use the object in algorithms only via computing a kernel function between a pair of objects. A kernel function is a similarity function satisfying certain properties. More precisely, a kernel function K over the object space X is binary function $K = X \times X \rightarrow [0, \infty]$ mapping a pair of objects $x, y \in X$ to their similarity score $K(x, y)$. A kernel function is required to be symmetric[1] and positive-semidefinite[2].

In this paper, we introduce a kernel-based method that uses clustering for word sense disambiguation. We define several kernels using information from a knowledge base for constructing the affinity matrix.

A. Common Ancestor Kernel

The goal of this kernel is to find a higher-level concept for summarizing a group of lower-level concepts so that concepts can be mapped to concepts of a coarse granularity. This kernel expresses that the similarity between each pair of concepts will be directly proportional to the number of common ancestors.

$$K_{common}(c, c') = |common\_ancestors(c, c')|$$

Where $common\_ancestors(c, c')$ is the set of all common ancestors of c and c'. That is,

$$common\_ancestors(c, c') = |\{c_x | c \leq c_x\} \cap \{c_y | c' \leq c_y\}|$$

B. Semantic Group Kernel

The semantic network has 132 semantic types and ensures a consistent categorization of all concepts represented in the Metathesaurus. The semantic types are classified into a smaller number of semantic groups. There are fifteen high-level semantic groups that help reduce the conceptual complexity of the large domain covered by the UMLS [19]. Groupings of semantic types - the semantic groups - may prove to be useful in a number of applications including improved visualization and display of the knowledge in a particular domain [20]; natural language processing, where higher level categories are sometimes sufficient for semantic processing [21]; and auditing a domain for the valid representation of concepts and their interrelationships [22]. Semantic groups allow classifying concepts and establishing dependencies between them at a higher level. To capture this knowledge a kernel is built, which relates two concepts if both belong to the same semantic group.

We can obtain a similarity matrix between concepts taking into account the semantic type of each one. Le S denote de concept-by-semantic_type matrix whose rows are indexed by the concepts and whose columns are indexed by semantic types. The $(i, j)th$ entry of S is 1 is the concept c_i has the

[1] "A binary function $K(\cdot, \cdot)$ is symmetric (over X), if $\forall x, y \in X, K(x, y) = K(y, x)$."

[2] "A binary function $K(\cdot, \cdot)$ is positive-semidefinitive, if $\forall x_1, x_2, \ldots, x_n \in X$ the $n \times n$ matrix $(K(x_i, x_j))_{ij}$ is positive-semidefinitive."

semantic type st_j; 0 otherwise. The matrix S gives rise to the similarity matrix $ST = SS^T$ between concepts.

For example, in the sentence *"Frozen shoulder due to cold damp treated with acupuncture and moxibustion on tender points."*, the concepts for *frozen* and *cold* in UMLS could be C0009264 (*An absence of warmth or heat or a temperature notably below an accustomed norm*) and C1550579 (*Keep frozen below* $0°C$) respectively. These concepts are not explicitly related but share the same semantic group: PHEN (Phenomena). With K_{sg} kernel this affinity is captured.

C. Relational Kernel

Concepts do not exist in isolation. They occur in complex, multidimensional networks that represent *"real world"* relationships. The primary link in the Semantic Network of UMLS is the *"isa"* relation. In addition, a set of non-hierarchical relations between the types has been identified. The following kernel accounts for the relations that are present on the KB:

$$K_{R_i}(c, c') = |\{R_i | \exists R_i(c, c') \in KB\}|$$

Where KB is composed of the set of UMLS concepts and the relations between them:

$$KB = \langle concepts, r : concepts \rightarrow concepts \rangle$$

D. Composite Kernel

We define the composite kernel as a linear combination of the individual kernels:

$$K(c, c') = K_{common}(c, c') + K_{sg}(c, c') + K_{R_i}(c, c')$$

Finally, for a set of concepts we can obtain a similarity matrix SM with the $(i, j)th$ entry equal to $K(c_i, c_j)$. This matrix is normalized as $S = D^{-1/2} SM D^{-1/2}$ in which D is a diagonal matrix with its (i, i)-element equal to the sum of the i-th row of SM. This is a usual normalization aimed at improving clustering tasks.

V. Clustering-based WSD Method

Our disambiguation method is an instance of the framework proposed in [23] . The underlying idea of the framework is to use clustering as a way of identifying semantically related word senses.

In this WSD method, the concepts are represented as profiles built from the repository of concepts of UMLS. A concept profile is a vector containing the words of the concept definition, or definitions, and its frequency that is normalized based on the inverted concept frequency similarly to the MRD method (see Section VI-B)

The disambiguation process starts from a clustering distribution of all possible concepts of the words. Then, clusters that match the best with the context are selected. If the selected clusters disambiguate the target word, the process stops and the concept of the word belonging to the selected clusters is interpreted as the disambiguating one. Otherwise, the clustering is performed again (regarding the remaining senses) until all words are disambiguated. Contexts and clusters are compared using the cosine distance similarly to the MRD approach presented below, where clusters are represented by their centroid of concepts profiles.

Concept clustering is carried out by the extended star clustering algorithm [24], which builds star-shaped and overlapped clusters. This clustering algorithm relies on a greedy cover of an affinity graph by star-shaped subgraphs [25]. The affinity graph is defined by the composite kernel defined in the previous section. Similar results are achieved with the bisecting k-means algorithm.

Notice that, all word concepts are included in the clustering process and the disambiguation is thus performed over all the concepts of all words in the sentence at once. The underlying hypothesis is that word concept clustering captures the reflected cohesion among the words of a sentence and each cluster reveals possible relationships existing among these word concepts. Thus, the way this clustering algorithm relates word concepts resembles the way in which syntactic and discourse relations link textual elements.

VI. Baseline knowledge-based WSD methods

We have compared the method presented above with available knowledge-based methods which are described in this section.

A. Journal Descriptor Indexing (JDI) Method

The JDI Method, introduced by [26], automatically assigns a concept to an ambiguous term by first identifying its semantic type with the assumption that each possible concept has a distinct semantic type. In this method, a semantic type vector is created for the semantic type of each of the possible concepts using one-word terms in the UMLS. A vector representing the ambiguous term is created using the words that exist in the same citation as the ambiguous term. The angle between this vector and each of the semantic type vectors is calculated using the cosine measure. The concept whose semantic type vector is closest to the vector representing the ambiguous word is assigned to the term. As this method relies on the semantic type(s) assigned to a concept, if two or more of the candidate concepts are assigned the same semantic type, this algorithm cannot disambiguate the ambiguous term.

The JDI experiments in this paper were conducted using the JDI implementation of this method and is available as part of the SPECIALIST Text Categorization tools [3].

[3] http://lexsrv3.nlm.nih.gov/Specialist/Summary/textCategorization.html

B. The Machine Readable Dictionary (MRD) Method

The MRD method uses context words surrounding the ambiguous word, which are compared to a profile built from each of the UMLS concept linked to the ambiguous term being disambiguated. Vectors of concept profiles linked to an ambiguous word and word contexts are compared using cosine similarity. The concept with the highest cosine similarity is selected. This method has been previously used by [27] in the biomedical domain.

A concept profile vector has as dimensions the tokens obtained from the concept definition, or definitions, if available, of synonyms and of related concepts (excluding siblings). Stop words are discarded , and Porter stemming is used to normalize the tokens. In addition, the token frequency is normalized based on the inverted concept frequency so that tokens which are repeated many times within the UMLS will have less relevance.

In order to perform disambiguation, the context of the ambiguous term is turned, as well, into a vector representation. The context vector for an ambiguous term includes the term frequency. The stop words are also removed, and the Porter Stemmer is applied. The word order, as in the concept profile, is lost in the conversion.

VII. Evaluation

Disambiguation methods are compared using the accuracy measure on a test set built on examples of MEDLINE citations with ambiguous words. The test set has been developed automatically using MeSH indexing from MEDLINE [28][4]. This set is based on the 2009AB version of the Metathesaurus and MEDLINE up to May 2010. The Metathesaurus is screened to identify ambiguous terms which contain MeSH headings. Then, each ambiguous term and the MeSH headings linked to it are used to recover MEDLINE citations using PUBMED where the term and only one of the MeSH headings co-occur. Because this initial set is noisy, we have filtered out some of the ambiguous terms to enhance precision of the set. This filtered set has 203 ambiguous words.

Table I shows the overall accuracy results of the MSH WSD data set. The data set is broken into three sections: the Abbreviation Set contains 106 ambiguous acronyms, the Term Set contains 88 ambiguous terms, and the Term/Abbreviation Set contains 9 ambiguous term/abbreviations. Since the JDI method is only able to disambiguate ambiguous terms or abbreviations whose possible senses do not share the same semantic type, there exist 44 ambiguous terms in which this method is not able to distinguish between the possible senses. From Table I we can conclude that althoug our method outperforms the JDI approach, it is outperformed by the MRD method.

Data set	JDI	MRD	Clustering-based
Abbreviation Set		0.8759	0.8157
Abbreviation Subset	0.6725	0.8838	0.8135
Term Set		0.7148	0.6548
Term Subset	0.6209	0.7132	0.6510
Term/Abbreviation Set		0.8801	0.7570
Term/Abbreviation Subset	0.6899	0.8715	0.7505
Overall Set		0.8070	0.7438
Overall Subset	0.6551	0.8118	0.7399

Table I
OVERALL ACCURACY ON THE DATA SET.

One possible reason of this is that we do not manage concept profiles during the clustering process (as the original of [23] did) mainly because of performance issues. Obtaining all the profiles for all the words appearing in the WSD collection similarly to MRD is very expensive and it implies a high overhead in the comparison of clusters and contexts. Instead we have used a BoW representation of the concept definitions (MRDEF) to build the cluster centroids for comparing with the context BoW representation.

Future work most focus on how to enrich the small profiles with similar contexts found in the same collection or other large KR like wikipedia.

VIII. Conclusion

In this paper we have presented our preliminary investigations on large scale disambiguation of semantically annotated biomedical corpora. First results show a good performance and an acceptable effectiveness, which outperform some state-of-the-art approaches. However, results must be improved in order to achieve the effectiveness of other methods like MRD. Future work will be focused on these improvements, mainly in the enrichment of context and concepts profiles, and in the addition of new kernels to account for more affinity-based features that can be useful to WSD.

References

[1] M. J. Schuemie, J. A. Kors, and B. Mons, "Word Sense Disambiguation in the Biomedical Domain: An Overview." *Journal of Computational Biology*, vol. 12, no. 5, pp. 554–565, 2005.

[2] T. Pedersen, "Unsupervised Corpus-Based Methods for WSD," in *Word Sense Disambiguation: Algorithms and Applications*, ser. Text, Speech and Language Technology. Dordrecht, The Netherlands: Springer, 2006, vol. 33, pp. 133–166.

[3] S. Brody and M. Lapata, "Bayesian word sense induction," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Morristown, NJ, USA: ACL, 2009, pp. 103–111.

[4] Available from: http://wsd.nlm.nih.gov/collaboration.shtml

[4] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice creamcone," in *Proceedings of the 5th annual international conference on Systems documentation*. ACM, 1986, pp. 24–26.

[5] F. Vasilescu, P. Langlais, and G. Lapalme, "Evaluating variants of the Lesk approach for disambiguating words," in *Proceedings of the Conference of Language Resources and Evaluations (LREC 2004)*, 2004, pp. 633–636.

[6] R. Navigli and M. Lapata, "An Experimental Study of Graph Connectivity for Unsupervised Word Sense Disambiguation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 4, pp. 678–692, 2010.

[7] P. Resnik, "Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language," *Journal of Artificial Intelligence Research*, vol. 11, no. 1, pp. 95–130, 1999.

[8] E. Agirre and G. Rigau, "Word sense disambiguation using Conceptual Density," in *Proceedings of the 16th conference on Computational linguistics - Volume 1*, ser. COLING '96. Stroudsburg, PA, USA: ACL, 1996, pp. 16–22.

[9] S. Banerjee, "Extended gloss overlaps as a measure of semantic relatedness," in *In Proceedings of the 18th International Joint Conference on Artificial Intelligence*, 2003, pp. 805–810.

[10] A. Budanitsky and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," *Computational Linguistics*, vol. 32, no. 1, pp. 13–47, mar 2006.

[11] R. Mihalcea, "Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, ser. HLT '05. Stroudsburg, PA, USA: ACL, 2005, pp. 411–418.

[12] R. Navigli and P. Velardi, "Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, pp. 1075–1086, July 2005.

[13] S. Brody, R. Navigli, and M. Lapata, "Ensemble methods for unsupervised WSD," in *ACL '06: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL*. Morristown, NJ, USA: ACL, 2006, pp. 97–104.

[14] E. Agirre and A. Soroa, "Personalizing PageRank for word sense disambiguation," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, ser. EACL '09. Stroudsburg, PA, USA: ACL, 2009, pp. 33–41.

[15] A. Jimeno-Yepes and A. R. Aronson, "Knowledge-based biomedical word sense disambiguation: comparison of approaches," *BMC Bioinformatics*, vol. 11, p. 569, 2010.

[16] O. Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology," *Nucl. Acids Res.*, vol. 32, no. suppl_1, pp. D267–270, jan 2004.

[17] T. Evgeniou, M. Pontil, and T. Poggio, "Statistical Learning Theory: A Primer," *International Journal of Computer Vision*, vol. 38, pp. 9–13, 2000.

[18] N. Cristianini and J. Shawe-Taylor, *An introduction to support Vector Machines: and other kernel-based learning methods*. New York, NY, USA: Cambridge University Press, 2000.

[19] A. McCray, A. Burgun, and O. Bodenreider, "Aggregating UMLS semantic types for reducing conceptual complexity." *Proceedings of Medinfo*, vol. 10, no. pt 1, pp. 216–20, 2001.

[20] M. S. T. Stuart J. Nelson, David D. Sherertz and M. S. Erlbaum, "Using MetaCard: a HyperCard browser for biomedical knowledge sources," *Proc Annu Symp Comput Appl Med Care*, pp. 151–154, 1990.

[21] T. C. Rindflesch and M. Fiszman, "The interaction of domain knowledge and linguistic structure in natural language processing: interpreting hypernymic propositions in biomedical text." *Journal of Biomedical Informatics*, vol. 36, no. 6, pp. 462–477, 2003.

[22] J. J. Cimino, "Auditing the Unified Medical Language System with semantic methods." *Journal of the American Medical Informatics Association : JAMIA*, vol. 5, no. 1, pp. 41–51, 1998.

[23] H. Anaya-Snchez, A. Pons-Porrata, and R. Berlanga-Llavori, "Word Sense Disambiguation Based on Word Sense Clustering," in *Advances in Artificial Intelligence - IBERAMIA-SBIA 2006*, ser. Lecture Notes in Computer Science, J. Sichman, H. Coelho, and S. Rezende, Eds. Springer, 2006, vol. 4140, pp. 472–481.

[24] A. Sanfeliu and J. Ruiz-Shulcloper, Eds., *Progress in Pattern Recognition, Speech and Image Analysis, 8th Iberoamerican Congress on Pattern Recognition, CIARP 2003, Havana, Cuba, November 26-29, 2003, Proceedings*, ser. Lecture Notes in Computer Science, vol. 2905. Springer, 2003.

[25] E. P. J. Aslam and D. Rus., "The Star Clustering Algorithm for Static and Dynamic Information Organization," *Journal of Graph Algorithms and Applications*, vol. 8, no. 1, pp. 95–129, 2004.

[26] S. M. Humphrey, W. J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. C. Rindflesch, "Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment," *Journal of the American Society for Information Science and Technology*, vol. 57, no. 1, pp. 96–113, 2006.

[27] B. T. McInnes, "An unsupervised vector approach to biomedical term disambiguation: integrating UMLS and Medline," in *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics on Human Language Technologies: Student Research Workshop*, ser. HLT-SRWS '08. Stroudsburg, PA, USA: ACL, 2008, pp. 49–54.

[28] A. Jimeno-Yepes, B. T. McInnes, and A. R. Aronson, "Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation," *BMC Bioinformatics*, 2011.