The BioWSD Project

Lexical Disambiguation for Biomedical Text

http://nlp.shef.ac.uk/BioWSD/

Robert Gaizauskas[+], Yikun Guo[+], David Martinez[*], Mark Stevenson[+]

[+] University of Sheffield, * University of Melbourne

Lexical Ambiguity in Medical Documents

In biomedicine the amount of published material has been growing exponentially in recent years, particularly in very productive areas, such as genomics. However, automatic processing of these documents is hampered by the fact that texts in the biomedical domain, like those in other areas, contain a range of lexical ambiguities. Studies have shown ~12% of the terms in MEDLINE citations are semantically ambiguous (may refer to more than one UMLS concept) (Weeber et. al., 2001).

BioWSD

The aim of the BioWSD project is to develop tools and algorithms to resolve various forms of lexical ambiguity found in biomedical texts. The resulting disambiguation systems will be integrated into Termino (Harkema et. al., 2005), a publicly available terminology recognition tool. In particular, the project will apply novel WSD methods to three distinct forms of lexical ambiguity:

1. **Terms which refer to multiple concepts**. For example, "cold" has six possible meanings in the UMLS Metathesaurus including "common cold", "cold sensation" and "Chronic Obstructive Airway Disease (COLD)".

2. **Ambiguous abbreviations**. Common in biomedical text (4.61 possible MEDLINE meanings on average). For example, “AC” has ten possible meanings including “atrioventricular connection”, “anterior colporrhaphy procedure” and “auditory cortex”.

3. **Terms with systematic relation between possible meanings** (regular polysemy). This form of ambiguity, which is particularly common in genomics literature, occurs when the same term can refer to a gene, protein or mRNA.

The NLM-WSD data set

50 highly ambiguous terms found in MEDLINE. 100 instances of each term were manually disambiguated by a group of 11 annotators. Average of 3.9 possible meanings in UMLS for each ambiguous term. (Weeber et. al., 2001)

All words

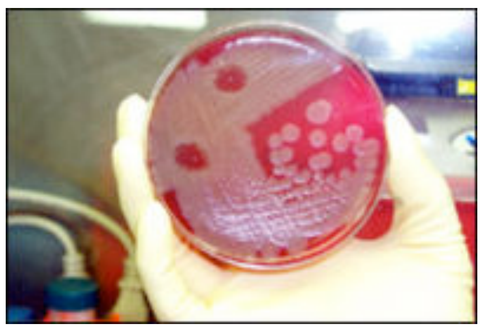
adjustment (*), association, blood pressure (*), cold (+), condition, culture, degree (+*), depression (+), determination, discharge (+), energy, evaluation (*), extraction (+), failure, fat (+), fit, fluid, frequency, ganglion, glucose, growth (+*), immunosuppression (*), implantation (+), inhibition, japanese (+), lead (+), man (+*), mole (+), mosaic (+*), nutrition (+*), pathology (+), pressure, radiation (*), reduction (+), repair (+*), resistance, scale (+*), secretion, sensitivity (*), sex (+), single, strains, support, surgery, transient, transport, ultrasound (+), variation, weight (+*), white (+*)

Subsets

Liu et. al. (2004) - 22 terms (denoted by +)
Leroy and Rindflesch (2005) - 15 terms (denoted by *)
Joshi et. al. (2005) - 28 terms (set union of Liu and Leroy subsets)
Common subset - 9 terms (intersection of Liu and Leroy subsets)

Example MEDLINE citations

“In peripheral blood mononuclear cell **culture** streptococcal erythrogenic toxins are able to stimulate tryptophan degradation in humans via the induction of interferon-gamma production.”
(culture = microorganism growth)



“Guidelines must be considered in light of local skills, **culture**, and resources, and need to be individualized to different patients and settings.”
(culture = society)

Lexical Disambiguation System

We use a supervised machine learning system that combines a variety of linguistic and domain-specific knowledge sources:

- **Linguistic features:** unigrams and lemmas from entire text, sentence and 8 word window; bigrams and trigrams around target word derived from lemmas, word forms and PoS tags; salient bigrams from text; syntactic dependencies
- **Concept Unique Identifiers (CUIs)** from UMLS assigned by MetaMap
- **Medical Subject Heading (MeSH) terms** assigned to MEDLINE abstract

Experiments have been carried out to compare a number of machine learning algorithms and feature sets. A Vector Space Model machine learning algorithm (Agirre and Martinez, 2004) was used for results reported here.

Results

Our best system outperforms previously published results on the NLM-WSD data set.

	Previous approaches			Our approach						
	Leroy & Rindflesch (2005)	Joshi et. al. (2005)	McInnes et. al. (2007)	Lingustic features	CUIs	MeSH terms	Ling + CUIs	**Ling + MeSH**	CUIs + MeSH	All features
All words			0.856	0.872	0.858	0.819	0.873	**0.878**	0.869	0.876
Joshi subset		0.825	0.800	0.823	0.796	0.766	0.824	**0.833**	0.814	0.826
Leroy subset	0.656	0.774	0.745	0.778	0.744	0.704	0.780	**0.790**	0.758	0.778
Liu subset		0.849	0.819	0.843	0.813	0.783	0.843	**0.851**	0.834	0.845
Common subset	0.688	0.798	0.756	0.796	0.751	0.704	0.796	**0.808**	0.769	0.792

Concept disambiguation in the medical domain benefits from the use of domain-specific features (UMLS CUIs and MeSH terms). Best results obtained using linguistic features in combination with MeSH terms.

References

Agirre and Martinez (2004) “The Basque Country Uninversity system: English and Basque tasks” *Proceedings of the workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*
Harkema, Roberts, Gaizauskas and Hepple (2005) “A Web Service for Biomedical Term Look-up” *Comparative and Functional Genomics 6(1-2), 86-93*
Joshi, Pedersen and Maclin (2005) "A Comparative Study of Support Vector Machines Applied to Supervised Word Sense Disambiguation Problems in the Medical Domain" *Proceedings of the 2nd Indian International Conference on Artificial Intelligence*
Leroy and Rindflesch (2005) "Effects of information and machine learning algorithms on word sense disambiguation with small data sets” *International Journal of Medical Informatics*
Liu, Teller and Friedman (2004) "A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation" *Journal of the American Medical Association*
McInnes, Pedersen and Carlis (2007) “Using UML Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain” *Proceedings of AMIA*
Weeber, Mork and Aronson (2001) “Developing a Test Collection for Biomedical Word Sense Disambiguation” *Proceedings of AAAI Symposium*

The University Of Sheffield.

nlp sheffield nlp

EPSRC
Engineering and Physical Sciences Research Council