Development of a Semantic Type Based WSD Tool

Chris J. Lu, Ph.D.[1], Susanne M. Humphrey [2], Willie J. Rogers[1], Allen C. Browne [2]
[1]Lockheed Martin/MSD, Bethesda, MD; [2]National Library of Medicine, Bethesda, MD (SMH retired)

Abstract

The Semantic Type Indexing (STI) tool, developed at the National Library of Medicine (NLM), is one of the most accurate unsupervised methods for the word sense disambiguation (WSD) applications. NLM's Lexical Systems Group (LSG) enhanced the STI tool, achieving an improved precision of 79.05% vs. baseline 73.81% when applied to NLM's WSD test collection. An ST-based WSD tool, STWSD, has been developed for distribution in the open source Text Categorization (TC) package.

1. Introduction

STI uses the context (phrase or sentence/s) to disambiguate an ambiguous word whose meanings represent different semantic types called ST candidates [1]. The ST candidate with the highest score/rank is presumed to be correct. ST Documents (ST-docs) and Journal Descriptor Indexing (JDI) scores are the two main elements of STI scores. An ST-doc is a set of one-word Metathesaurus strings associated with an ST. As previously reported, JDI is a sophisticated methodology with consistent results [2] for categorizing input text according to biomedical specialties, known as JDs. An optimal ST-doc contains words which best represent the ST; the better the representation, the better the STI result. In this study, we enhanced ST-docs and WSD algorithms to improve the precision of WSD. Standard STI uses pre-computed word-ST vectors derived from the similarity between JDI of words and JDI of ST-docs. Therefore STI does not require actually running JDI. However, a process known as ST real-time indexing (STRI) results in the ST indexing of text by actually running JDI on input text and comparing the JDI of this text to the JDI of the ST-docs. STRI was used for refining ST-docs, resulting in improved precision, which subsequently carried over to better precision by standard STI based on word-ST vectors.

2. ST Documents Enhancement

We developed a test suite to test precision of all 100 (instead of 67 in [1]) instances for each of the 45 ambiguous words in NLM's WSD test collection. The TC package contains two releases – 2007 and 2008, each with different ST-docs and different sets of words in the JDI training set. The 2008 release uses a more recent MEDLINE training set and was determined to be reliable in [2] and is preferred. Thus, our baseline precision is that of the 2008 release, 73.81% (Table 1 row A). Various WSD tests on all four permutations of the two ST-docs releases and two JDI releases showed that precision of the 2008 release was worse than that of 2007, and the likely cause was inferior ST-docs in the 2008 release. We then proceeded to modify them in order to improve precision.

First, instead of counting a word only once in ST-docs, we used occurrence data in the form of weighted frequency (WF). Precision was improved to 76.29% (Table 1 row B). Second, a word can be associated with multiple concepts (CUIs) and thus associated with multiple STs and ST Groups (SGs). Such words are ambiguous and usually not good representatives for associated STs. Accordingly, we tested ST-docs with words associated with only one SG (1SG). Precision improved to 76.85% (Table 1 row C).

Theoretically, a good representative word for an associated ST in an ST-doc should have a high score for that ST. STRI was used as a filter to refine an ST-doc with two rules: a good representative word a) must be ranked in the top n (Top n) where n is to be determined, or b) the ST score must be within one standard deviation (StdDev) of the top score. We applied several STRI filters on WF-1SG ST-docs and found the best precision of 78.07% with StdDev & Top15 (Table 1 row D). Finally, we added back words associated with multiple SGs (MSG) if the ST was in the top 3. Precision improved to 78.71% (Table 1 row E).

ID	ST Documents	Prec.
A	Baseline – 2008	73.81%
B	Weighted Frequency (WF)	76.29%
C	WF-1SG	76.85%
D	WF-1SG: StdDev & Top15	78.07%
E	WF-1SG: StdDev & Top15; MSG: Top3	78.71%
F	ST Documents E with CS	79.05%

Table 1. Improved WSD precision in STWSD tool

3. WSD Tool Enhancement

STI uses two independent methods to score an ST: one based on word count (WC), the other on document count (DC) for the word; precision in Table 1 Rows A-E is DC based. We combined the scores (CS) by selecting the ST with the higher score, after computing the score of each candidate ST as the absolute difference between the WC and DC based scores. Precision was further improved to 79.05% (Table 1 row F). Several key features were implemented for this WSD tool, such as Java APIs and stand alone tools; the ambiguous sentences option [1] to get better precision when the context of the text input is sentences; forcing the input ambiguous word and its morphological variants to be legal words (e.g., by temporarily removing them from our stopword list) to avoid an empty result.

4. Conclusion

We enhanced ST-docs and WSD algorithms to reach precision of 79.05% on NLM's WSD test collection. Based on this study, LGS plans to distribute the STWSD tool in the open source TC package.

References

1. Humphrey SM, et al. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: preliminary experiment. J Am Soc Inf Sci Technol 2006, 57(1):96-113. Erratum in: J Am Soc Inf Sci Technol 2006 Mar;57(4):726.
2. Lu CJ, et al. A method for verifying a vector-based text classification system. AMIA 2008, 1030-31.