Automatic Sense Disambiguation for Acronyms

Manuel Zahariev
Amware Enterprises Ltd.,
PO Box 3674, Garibaldi Highlands, B.C., V0N 1T0 Canada
manuel@amware.com

ABSTRACT

A machine learning methodology for the disambiguation of acronym senses is presented, which starts from an acronym sense dictionary. Training data is automatically extracted from downloaded documents identified from the results of search engine queries. Leave-one-out cross-validation on 9,963 documents with 47 acronym forms achieves accuracy 92.58% and $F_{\beta=1}$=91.52%.

Categories and Subject Descriptors: H.3.1Content Analysis and Indexing: Linguistic processing

General Terms: Languages.

Keywords: acronyms, abbreviations.

1. INTRODUCTION

Acronyms are a systematic form of abbreviation, and a significant and arguably the most dynamic portion of the lexicon of many languages. They have been studied mainly in English, [5].

In spite of well known writing style requirements stating that all acronyms need to be explicitly defined (at their first occurrence) in every document where used, naturally occurring text often uses acronyms which are assumed to be well known in the domain. This can create serious understanding difficulties for non-expert readers, as well as for the automated processing (information extraction, automated understanding, etc.) of text.

Acronym sense disambiguation is the problem of identifying the sense (or *expansion*) associated with a given acronym form (or *spelling*) occurring in text. This has been studied for acronyms in the medical domain [3] (89% accuracy) using supervised machine learning but has so far been an open problem for general text.

The polysemy (property to have multiple senses) of acronyms is more productive than that of regular words. In a 2001 version of the WWWAAS (World-Wide Web Acronym and Abbreviation Server) database[1] containing (after cleanup) 16,823 terms, only 52.03% of acronym forms have only one sense, compared to 81.72% of 136,972 term senses in WordNet [1].

Given acronym dictionaries of adequate coverage, acronym sense disambiguation represents a special case of the more general problem of Word Sense Disambiguation (WSD), one of the most difficult and elusive open problems in Natural Language Processing. The main difficulties of WSD lie in the fluid definition of *word sense* and the high costs of acquiring consistent sense repositories, of creating training sets of adequate coverage, and of conducting unbiased large-scale evaluations. It is proposed here that general WSD difficulties either do not exist at all for acronym sense disambiguation, or can be surmounted through methods shown further, using data and resources readily available on the Internet.

[1]I would like to thank Peter Flynn from the University College of Cork, Ireland for providing the WWWAAS database for the purposes of this research.

SIGIR'04, July 25–29, 2004, Sheffield, South Yorkshire, UK.
ACM 1-58113-881-4/04/0007.

2. SYSTEM

A machine learning system is proposed, using support vector machines (SVM) in a pattern recognition configuration, with a linear kernel, in the implementation of Joachims [2]. Features are terms occurring in the same document as the target acronym form. Terms are either words that occur in WordNet or other acronym forms. The value of each feature ($\delta_D(t, A)$) is used to model both the presence of a given term (t) in the same document (D) as the target acronym form (A) and the distance in words (d) between occurrences of the term (t_i) and occurrences of the target acronym (A_j):

$$\delta_D(t, A) = \begin{cases} \alpha + \dfrac{\beta}{\min\limits_{t_i, A_j \in D} d(t_i, A_j)} & \text{if} \quad t \in D \\ 0 & \text{if} \quad t \notin D \end{cases}$$

The system-wide constants $\alpha = 0.25$ and $\beta = 1.0$ are chosen so that the presence of a given term in the same document as the target acronym accounts quantitatively as much as the proximity within a four-word window around each target acronym occurrence. The impact of common words is reduced by requiring a minimum length for WordNet word features. This is intuitively adequate, following Zipf's law (shorter words are fewer and occur more frequently than longer words).

For each document-acronym form pair, the values for $\delta_D(t, A)$ for all terms $t \in D$ are calculated and collected into one feature vector. For each acronym sense, the system is trained on collections of documents (value vectors) with known senses. Positive examples are documents (and associated value vectors) containing occurrences of the acronym form with the target sense, and negative examples are value vectors corresponding to documents where the acronym form occurs with other senses.

Positive emphasis is calculated for each acronym sense (S_A) from the number of documents, $N(S_A)$, containing the acronym form A with the sense S_A and the number of documents, $N(\overline{S_A})$, containing other senses of the acronym form A :

$$\epsilon^+(S_A) = \left\lceil \frac{w^+ \times N(\overline{S_A})}{N(S_A)} \right\rceil$$

The value $w^+ = 2$ is a system constant, chosen to favor false positive errors compared to false negative errors (increase recall at the expense of precision).

The result of training is a set of *classification models*, one for each sense of a given acronym. Feature vectors are calculated for documents with occurrences of acronym forms with unknown senses. Disambiguation is performed by two tasks: *decision* (a boolean answer to the question whether a given acronym occurrence is associated with a given sense) and *selection* (the choice of one of a list of senses for a given acronym form occurrence).

The decision task is based on the result of "pattern recognition" using the learned classification model for the acronym sense, applied to the target document's feature vector.

The selection task identifies the maximum value of the decision function resulting from the decision task applied repeatedly to the value vector of the target document, using consecutively the classification models for each sense.

To illustrate, consider an acronym form X, with n senses: $s_1..s_n$. The training set for senses of the acronym form X is a set of value vectors, V_X. For each sense s_k ($k \in 1..n$), separately, all vectors $u \in V_X$, with Sense(u) $= s_k$ are considered positive examples (labeled $+1$). All other vectors in V_X are considered negative examples (labeled -1). A classification model for s_k is obtained through SVM training on V_X such labeled, and is used to classify a new vector v, resulting in a classification value $c(s_k, v)$. For each sense, the result of the classification decision is given by:

- Sense(v) $= s_k$ if $c(s_k, v) \geq 0$
- Sense(v) $\neq s_k$ if $c(s_k, v) < 0$

The result of selection is given by:

$$\text{Sense}(v) = s_k \text{ where } k = \arg_k \max_{1 \leq k \leq n} c(s_k, v)$$

It is expected that classification on only one of the senses will return a positive classification value. While the feature spaces of the classification models for different senses are unrelated, in situations of ambiguity (where zero or more than one sense are returned as accurate), we chose as adequate the sense resulting in minimum distance from the separation hyperplane represented by the classification model.

3. ACQUISITION OF TRAINING DATA

The heuristic of *one acronym sense per document*, similar with the *one word sense per discourse* heuristic [4] is used to collect automatically training data from the Internet. Exceptions are represented by documents which are, or contain, wide coverage lists of acronyms. Those are removed by requiring that each document taken into account contains at most a given number of acronym forms, a system constant ($\gamma_A = 100$). A limitation of the size in tokens of each document (a system constant γ_t = 10,000) is also imposed, based on the assumption that very large documents (e.g. whole books) may induce noisy input during training.

The cooccurrence in a given document of an acronym form and the complete phrase of a corresponding acronym expansion (sense) is taken to indicate the sense of the acronym form to be the acronym expansion. For example, the cooccurrence in a document of the acronym form 'SPS' and one of its expansions "solar power satellite" indicates the sense "solar power satellite" for all occurrences of 'SPS' within the document.

Documents containing both an acronym and its expansion are located through search engine (such as Google) queries for both the acronym form and the full phrase of its expansion. Web documents in the Adobe PDF format are used for training, as they are considered more likely sources of natural occurrences of specific acronyms in text than HTML documents. For example, the following query to Google will return links to PDF documents which contain occurrences of the acronym form SPS and its expansion "solar power satellite":

SPS "solar power satellite" filetype:pdf

For each acronym-expansion pair, 100 documents (or at least as many as available in the result of the query) are downloaded. Each document is converted to text. In each of the resulting (successfully converted) text documents, all occurrences of the expansion are replaced with the acronym form (in the example above all occurrences of "solar power satellite" are replaced with 'SPS') obtaining documents where the target acronym occurs with a given sense, and without an explicit indication of its expansion. Each document is marked with the specific acronym sense.

The result of the training data acquisition phase is a training corpus, containing a training set for each acronym, with documents containing acronym forms, but not explicit acronym expansions, and labeled using the senses (expansions) of the acronym. The training corpus consists of 9,963 sense-marked documents (feature-value vectors), distributed between 167 senses of 47 acronyms, and is built starting from acronym forms with more than three senses randomly chose from the WWWAAS database.

4. EVALUATION AND CONCLUSIONS

Leave-one-out cross validation for each sense of each acronym in the training corpus is performed. The results are consolidated in global performance decision and selection figures: $F_{\beta=1}$ = 91.52% (precision: 90.57%, recall: 92.48%) and accuracy 92.58%. Baseline $F_{\beta=1}$ and accuracy are calculated for always choosing the most frequent sense of each given acronym form and are both at 36.94%. The performance of the system is lowest for acronym senses for which only a small number of documents can be identified, downloaded and converted to feature-value vectors, usually due to overtraining of the better-represented sense. For example, DEC is correctly disambiguated as "device clear" with a recall of only 38.89%, using only 18 documents as evidence (of a total of 164 for three senses). Senses of acronyms which are semantically equivalent (*false polysemy*) also account for reduced performance, such as the equivalent expansions "end of text" ($F_{\beta=1} = 63.67\%$) and "end of transmission" ($F_{\beta=1} = 71.30\%$) of 'EOT'. Detailed results of the evaluation are presented in [5]. System performance is superior to that of general-purpose WSD systems.

The automatic acquisition of training data, starting from dictionaries of acronyms and using documents available on the Internet, returned from search engine queries, is a low cost alternative to the major expense of building WSD training corpora. Automatic leave-one-out cross validation on training data reduces the cost and the bias involved in building evaluation corpora. The performance of the system on given acronyms and senses, as calculated during evaluation, can also be used in runtime conditions to qualify the confidence in selection of a given sense on a new document.

5. REFERENCES

[1] Christiane Fellbaum, editor. *WordNet, An Electronic Lexical Database*. MIT Press, 1998.

[2] Thorsten Joachims. Making large-scale svm learning practical. In B. Schölkopf, Christopher J.C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT-Press, 1999.

[3] Serguei Pakhomov. Semi-supervised maximum entropy based approach to acronym and abbreviation normalization in medical texts. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) Philadelphia, July 2002*, 2002.

[4] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 1995.

[5] Manuel Zahariev. *A (Acronyms)*. PhD thesis, Simon Fraser University, School of Computing Science, 2004.