Combined One Sense Disambiguation of Abbreviations

Yaakov HaCohen-Kerner
Department of Computer Science, Jerusalem College of Technology (Machon Lev)
21 Havaad Haleumi St., P.O.B. 16031, 91160 Jerusalem, Israel
kerner@jct.ac.il

Ariel Kass
Department of Computer Science, Jerusalem College of Technology (Machon Lev)
21 Havaad Haleumi St., P.O.B. 16031, 91160 Jerusalem, Israel
ariel.kass@gmail.com

Ariel Peretz
Department of Computer Science, Jerusalem College of Technology (Machon Lev)
21 Havaad Haleumi St., P.O.B. 16031, 91160 Jerusalem, Israel
relperetz@gmail.com

Abstract

A process that attempts to solve abbreviation ambiguity is presented. Various context-related features and statistical features have been explored. Almost all features are domain independent and language independent. The application domain is Jewish Law documents written in Hebrew. Such documents are known to be rich in ambiguous abbreviations. Various implementations of the one sense per discourse hypothesis are used, improving the features with new variants. An accuracy of 96.09% has been achieved by SVM.

1 Introduction

An abbreviation is a letter or sequence of letters, which is a shortened form of a word or a sequence of words, which is called the sense of the abbreviation. Abbreviation disambiguation means to choose the correct sense for a specific context.

Jewish Law documents written in Hebrew are known to be rich in ambiguous abbreviations (HaCohen-Kerner et al., 2004). They can, therefore, serve as an excellent test-bed for the development of models for abbreviation disambiguation.

As opposed to the documents investigated in previous systems, Jewish Law documents usually do not contain the sense of the abbreviations in the same discourse. Therefore, the abbreviations are regarded as more difficult to disambiguate.

This research defines features, as well as experiments with various variants of the one sense per discourse hypothesis. The developed process considers other languages and does not define pre-execution assumptions. The only limitation to this process is the input itself: the languages of the different text documents and the man-made solution database inputted during the learning process limit the datasets of documents that may be solved by the resulting disambiguation system.

The proposed system, preserves its portability between languages and domains because it does not use any natural language processing (NLP) sub-system (e.g.: tokenizer and tagger). In this matter, the system is not limited to any specific language or dataset. The system is only limited by the different inputs used during the system's learning stage and the set of abbreviations defined.

This paper is organized as follows: Section 2 presents previous systems dealing with disambiguation of abbreviations. Section 3 describes the features for disambiguation of Hebrew abbreviations. Section 4 presents the implementation of the one sense per discourse hypothesis. Section 5 describes the experiments that have been carried out. Section 6 concludes and proposes future directions for research.

2 Abbreviation Disambiguation

The one sense per collocation hypothesis was introduced by Yarowsky (1993). This hypothesis states that natural languages tend to use consistent spoken and written styles. Based on this hypothesis, many terms repeat themselves with the same meaning in all their occurrences. Within the context of determining the sense of an abbreviation, it may be assumed that authors tend to use the same words in the vicinity of a specific long form of an abbreviation. The words may be reused as indicators of the proper solution of an additional unknown abbreviation with the same words in its vicinity. This is the basis for all contextual features defined in this research.

The one sense per discourse hypothesis (OS) was introduced by Gale et al. (1992). This

Proceedings of ACL-08: HLT, Short Papers (Companion Volume), pages 61–64,
Columbus, Ohio, USA, June 2008.

hypothesis assumes that in natural languages, there is a tendency for an author to be consistent in the same discourse or article. That is, if in a specific discourse, an ambiguous phrase or term has a specific meaning, any other subsequent instance of this phrase or term will have the same specific meaning. Within the context of determining the sense of an abbreviation, it may be assumed that authors tend to use a specific abbreviation in a specific sense throughout the discourse or article.

Research has been done within this domain, mainly for English medical documents. Systems developed by Pakhomov (2002; 2005), Yu et al. (2003) and Gaudan et al. (2005) achieved 84% to 98% accuracy. These systems used various machine learning (ML) methods, e.g.: Maximum Entropy, SVM and C5.0.

In our previous research (HaCohen-Kerner et al., 2004), we developed a prototype abbreviation disambiguation system for Jewish Law documents written in Hebrew, without using any ML method. The system integrated six basic features: common words, prefixes, suffixes, two statistical features and a Hebrew specific feature. It achieved about 60% accuracy while solving 50 abbreviations with an average of 2.3 different senses in the dataset.

3 Abbreviation Disambiguation Features

Eighteen different features of any abbreviation instance were defined. They are divided into three distinct groups, as follows:

Statistical attributes: Writer/Dataset Common Rule (WC/DS). The most common solution used for the specific abbreviation by the discussed writer/ in the entire dataset.

Hebrew specific attribute: Gimatria Rule (GM). The numerical sum of the numerical values attributed to the Hebrew letters forming the abbreviation (HaCohen-Kerner et al., 2004).

Contextual relationship attributes:

1. **Prefix Counted Rule (PRC)**: The selected sense is the most commonly appended sense by the specific prefix.
2. **Before/After K (1,2,3,4) Words Counted Rule (BKWC/AKWC):** The selected sense is the most commonly preceded/succeeded sense by the K specific words in the sentence of the specific abbreviation instance.
3. **Before/After Sentence Counted Rule (BSC/ASC):** The selected sense is the most commonly preceded/succeeded sense by all the specific words in the sentence of the specific abbreviation instance.
4. **All Sentence/Article Counted Rule (AllSC/AllAC):** The selected sense is the most commonly surrounded sense by all the specific words in the sentence/article of the specific abbreviation instance.
5. **Before/After Article Counted Rule (BAC/AAC):** The selected sense is the most commonly preceded/succeeded sense by all the specific words in the article of the specific abbreviation instance.

4 Implementing the OS Hypothesis

As mentioned above, the basic assumption of the OS hypothesis is that there exists at least one solvable abbreviation in the discourse and that the sense of that abbreviation is the same for all the instances of this abbreviation in the discourse. The correctness of all the features was investigated based on this hypothesis for several variants of "one sense" based on the discussed discourse: none (No OS), a sentence (osS), an article (osA) or all the articles of the writer (osW).

The OS hypothesis was implemented in two forms. The “pure” form (with the suffix S/A/W without C) uses the sense found by the majority voting method for an abbreviation in the discourse and applies it “blindly” to all other instances.

The “combined” form (with the suffix C) tries to find the sense of the abbreviation using the discussed feature only. If the feature is unsuccessful, then we use the relevant one sense variant using the majority voting method. This form is derived from the possibility that more than one sense may be used within a single discourse and only instances with an unknown sense conform to the hypothesis.

The use of the OS hypothesis, in both forms, is only relevant for context based features, since the solutions by other features are static and identical from one instance to another.

Therefore, for each of the 15 context based features, 6 variants of the hypothesis were implemented. This produces 90 variants, which together with the 18 features in their normal form, results in a total of 108 variants. In addition, the ML methods were experimented together with the OS hypothesis. Of the 108 possible variants, for the 18 features, the best variant for each feature

was chosen. In each step, the next best variant is added, starting from the 2 best variants.

5 Experiments

The examined dataset includes Jewish Law Documents written by two Jewish scholars: Rabbi Y. M. HaCohen (1995) and Rabbi O. Yosef (1977; 1986). This dataset includes 564,554 words where 114,814 of them are abbreviations instances, and 42,687 of them are ambiguous. That is, about 7.5% of the words are ambiguous abbreviations. These ambiguous abbreviations are instances of a set of 135 different abbreviations. Each one of the abbreviations has between 2 to 8 relevant possible senses. The average number of senses for an abbreviation in the dataset is 3.27.

To determine the accuracy of the system, all the instances of the ambiguous abbreviations were solved beforehand. Some of them were based on published solutions (HaCohen, 1995) and some of them were solved by experienced readers.

5.1 Results of the variants of OS Hypothesis

The results of the OS hypothesis variants, for all the features, are presented in Table 1. These results are obtained without using any ML methods.

Use of OS / Feature	Accuracy Percentage %						
	No OS	osS	osSC	osA	osAC	osW	osWC
PRC	33.67	34.41	34.52	52.77	54.54	66.66	71.04
B1WC	56.05	56.41	56.61	67.74	71.84	72.93	82.51
B2WC	55.72	56.23	56.35	69	72.34	74.85	82.84
B3WC	60.54	60.89	61.01	72.67	75.48	75.44	82.86
B4WC	64.49	64.72	64.85	74.29	76.5	75.52	82.2
BSC	75.21	75.18	75.24	76.85	78.15	74.92	78.52
BAC	76	76	76	76.01	76	75.39	76
A1WC	**78.79**	79.01	79.21	78.72	83.81	76.32	**87.75**
A2WC	77.57	78.07	78.26	79.15	83.43	78.54	87.62
A3WC	78.64	79.11	79.28	79.61	83	78.19	85.8
A4WC	75.44	79.28	79.5	79.41	82.42	78.01	84.99
ASC	78.59	78.61	78.62	78.25	78.94	77.37	79.04
AAC	75.44	75.44	75.44	75.34	75.44	77.28	75.44
AllSC	77.97	77.97	77.97	77.9	78.02	77.22	78.04
AllAC	74.12	74.12	74.12	74.12	74.12	76.93	74.12
GM	**46.82**	46.82	46.82	46.82	46.82	46.82	**46.82**
WC	**82.84**	82.84	82.84	82.84	82.84	82.84	**82.84**
DC	**78.34**	78.34	78.34	78.34	78.34	78.34	**78.34**

Table 1. Results of the OS Variants for all the Features.

The two best pure features were WC and A1WC with 82.84% and 78.79% of accuracy, respectively. The first finding shows that about 83% of the abbreviations have the same sense in the whole dataset. The second finding shows that about 79% of the abbreviations can be solved by the first word that comes after the abbreviation.

Generally, contextual features based on the context that comes after the abbreviation, achieve considerably better results than all other contextual features. Specifically, the A1WC_osWC feature variant achieves the best result with 87.75% accuracy. These results suggest that each individual abbreviation has stronger relationship to the words after a specific instance, especially to the first word.

Almost every feature has at least one variant that achieves a substantial improvement in results compared the results achieved by the feature in its normal form. The average relative improvement is about 18%.

For all features, except BAC, the best variant uses the OS implementation with the discourse defined as the entire dataset. This may be attributed to the similarity of the different articles in the dataset. This is supported by the fact that the best feature, in its normal form, is the WC feature.

In addition, for all but three features (BAC, AAC, AllAC), the best variant used the combined form of the OS implementation. This is intuitively understandable, since "blindly" overwriting probably erases many successes of the feature in its normal form.

5.2 The Results of the Supervised ML Methods

Several well-known supervised ML methods have been selected: artificial neural networks (ANN), Naïve Bayes (NB), Support Vector Machines (SVM) and J48 (Witten and Frank, 1999) an improved variant of the C4.5 decision tree induction. These methods have been applied with default values and no feature normalization using Weka (Witten and Frank, 1999). Tuning is left for future research. To test the accuracy of the models, 10-fold cross-validation was used.

Table 2 presents the results of these supervised ML methods, by incrementally combining the best variant for each feature (according to Table 1).

Table 2 shows that SVM achieved the best result with 96.09% accuracy. The best improvement is

about 13%, from 82.84% accuracy for the best variant of any feature to 96.02% accuracy. This table also reveals that incremental combining of most of the variants leads to better results for most of the ML methods.

# of Vari-ants	Variants / ML Method	ANN	NB	SVM	J48
2	**A1WC_osWC +A2WC_osWC**	91.56	91.40	94.29	91.94
3	**+ A3WC_osWC**	91.72	91.42	94.43	92.20
4	**+ A4WC_osWC**	91.75	91.51	94.43	92.34
5	**+ B3WC_osWC**	92.68	92.11	95.33	93.33
6	**+ WC**	92.95	92.16	95.71	93.54
7	**+ B2WC_osWC**	92.81	91.79	95.67	93.59
8	**+ B1WC_osWC**	92.91	91.06	95.68	93.56
9	**+ B4WC_osWC**	92.83	91.15	95.62	93.55
10	**+ ASC_osWC**	92.83	91.10	95.60	93.52
11	**+ BSC_osWC**	92.95	91.17	95.65	93.58
12	**+ DC**	92.98	91.17	95.63	93.58
13	**+ AllSC_osWC**	92.82	91.50	95.63	93.58
14	**+ AAC_osW**	92.84	91.42	95.59	93.58
15	**+ AllAC_osW**	93.10	91.43	95.77	93.58
16	**+ BAC_osA**	93.09	91.28	95.79	93.70
17	**+ PRC_osWC**	93.25	91.50	**96.09**	93.71
18	**+ GM**	93.28	91.52	96.02	93.93

Table 2. The Results of the ML Methods.

The comparison of the SVM results to the results of previous (Section 2) shows that our system achieves relatively high accuracy. However, most previous systems researched ambiguous abbreviations in the English language, as well as different abbreviations and texts.

6 Conclusions, Summary and Future Work

This is the first ML system for disambiguation of abbreviations in Hebrew. High accuracy percentages were achieved, with improvement ascribed to the use of OS hypothesis combined with ML methods. These results were achieved without the use of any NLP features. Therefore, the developed system is adjustable to any specific type of texts, simply by changing the database of texts and abbreviations.

This system is the first that applies many versions of the one sense per discourse hypothesis. In addition, we performed a comparison between the achievements of four different standard ML methods, to the goal of achieving the best results, as opposed to the other systems that mainly focused on one ML method, each.

Future research directions are: comparison to abbreviation disambiguation using the standard bag-of-words or collocation feature representations, definition and implementation of other NLP-based features and use of these features interlaced with the already defined features, applying additional ML methods, and augmenting the databases with articles from additional datasets in the Hebrew language and other languages.

References

Y. M. Hacohen (Kagan). 1995. *Mishnah Berurah* (in Hebrew), Hotzaat Leshem, Jerusalem.

Y. HaCohen-Kerner, A. Kass and A. Peretz. 2004. *Baseline Methods for Automatic Disambiguation of Abbreviations in Jewish Law Documents*. Proceedings of the 4th International Conference on Advances in Natural Language LNAI, Springer Berlin/Heidelberg, 3230: 58-69.

W. Gale, K. Church and D. Yarowsky. 1992. *One Sense Per Discourse*. Proceedings of the 4th DARPA speech in Natural Language Workshop, 233-237.

S. Gaudan, H. Kirsch and D. Rebholz-Schuhmann. 2005. *Resolving abbreviations to their senses in Medline*. Bioinformatics, 21 (18): 3658-3664.

S. Pakhomov. 2002. *Semi-Supervised Maximum Entropy Based Approach to Acronym and Abbreviation Normalization in Medical Texts*. Association for Computational Linguistics (ACL), 160-167.

S. Pakhomov, T. Pedersen and C. G. Chute. 2005. *Abbreviation and Acronym Disambiguation in Clinical Discourse*. American Medical Informatics Association Annual Symposium, 589-593.

D. Yarowsky. 1993. *One sense per collocation*. Proceedings of the workshop on Human Language Technology, 266-271.

O. Yosef. 1977. *Yechave Daat* (in Hebrew), Publisher: Chazon Ovadia, Jerusalem.

O. Yosef. 1986. *Yabia Omer* (in Hebrew), Publisher: Chazon Ovadia, Jerusalem.

Z. Yu, Y. Tsuruoka and J. Tsujii. 2003. *Automatic Resolution of Ambiguous Abbreviations in Biomedical Texts using SVM and One Sense Per Discourse Hypothesis*. SIGIR'03 Workshop on Text Analysis and Search for Bioinformatics.

H. Witten and E. Frank. 2007. *Weka 3.4.12: Machine Learning Software in Java:* http://www.cs.waikato.ac.nz/~ml/weka.