Research Progress of Polysemous Disambiguation

Chao Han[a], Ha Si[b]

College of Computer Science and Technology Inner Mongolia Normal University Hohhot, China

[a]chaohan950101@126.com, [b]hasi76@126.com

Abstract. Linguistic ambiguity is a common phenomenon in human language communication. In natural language, when a word in a sentence has more than one meaning, ambiguity may occur. Word sense disambiguation is an important research topic in the field of natural language processing. The disambiguation effect has a significant impact on machine translation, information retrieval, information extraction and text mining, speech recognition and other aspects, so word meaning disambiguation has important theoretical research and practical application significance. The author analyzes, compares, and summarizes Knowledge-based Methods, Supervised Methods, and Unsupervised Methods, and analyzes the future trends of word sense disambiguation.

Keywords: Mongolian, Natural language processing, Word sense disambiguation, Semantic category.

1. Introduction

Word Sense Disambiguation [1]is the core of natural language processing tasks. A word with multiple meanings has different specific meanings in different contexts [2] For example, in Mongolian "ᠬᠠᠮᠤᠭ", This word has multiple meanings depending on the context and part of speech. When it is used as an adjective, it has the following meanings: 1 black, such as "ᠬᠠᠮᠤᠭ ᠪᠤᠤ" (black cloth); 2 violent, strong, intense, such as "ᠬᠠᠮᠤᠭ ᠰᠦᠯᠬᠢᠨ" (storm); 3 heavy, stupid: such as "ᠬᠠᠮᠤᠭ ᠠᠵᠢᠯ" (coarse work); 4 plain, blank; 5 fierce, vicious, insidious; 6 secular; And when it is used as a noun, the meaning is sinful and toxic. For polysemy in sentences, the purpose of word sense disambiguation is to determine the specific semantics of words in the context, otherwise it will directly affect people's understanding of the whole context or even the whole article. It can be said that the use of computers for natural language processing is a process of constant disambiguation. The problem of polysemy disambiguation is the key to the field of information processing, which directly affects the research results including text classification, machine translation, and speech recognition in the field of natural language processing.

In summary, there are three main methods for word sense disambiguation [3]: Knowledge-based Methods, Supervised Methods, Unsupervised Methods. Knowledge-based Methods uses artificial dictionaries or dictionaries to obtain the relationship between semantics of words to disambiguate, the accuracy of such methods is generally higher than the corpus-based approach, but the prerequisite for using this method is to have a large semantic knowledge base such as Wordnet [4] and Hownet [5]. Supervised Methods uses artificially pre-labeled word meaning information for word sense disambiguation. Unsupervised Methods uses a corpus that has not been manually annotated for disambiguation and is disambiguated by machine learning. Among these three methods, the Supervised Methods disambiguation effect is better than the other two methods, but since the corpus is manually labeled by linguistic experts, it often requires a lot of human and material support.

2. Knowledge-based Methods

Knowledge-based Methods are a relatively popular research method [6]. In the study of numerous word sense disambiguation methods, except for unsupervised machine learning studies that do not require the support of dictionary resources, Other various word sense disambiguation methods almost all require a dictionary to assist. Both the machine-readable dictionary and the semantic dictionary provide rich semantic knowledge for vocabulary, which is the main source of knowledge acquisition for word sense disambiguation. Knowledge-based methods mainly refer to the acquisition of semantic knowledge and the relationship between semantics through the construction of a large-scale

knowledge base to assist the disambiguation. It uses the dictionary as a knowledge base for disambiguation to reduce the dependence of the disambiguation system on the size of training corpus data and artificially labeled corpus. The knowledge-based methods mainly include a machine-readable dictionary-based method and a semantic-based dictionary-based method.

2.1 Machine-readable Dictionary Method

The machine-readable dictionary was compiled by linguists who spent a lot of time and effort. Calculate the different meanings of vocabulary and the co-occurrence rate of adjacent words of ambiguous words, First select the ambiguous words, and then extract the adjacent words of the ambiguous words, Use the machine to find the different meanings of this ambiguity in the machine readable dictionary, and find out the meaning of the word adjacent to the ambiguous word, Conduct detailed analysis and select the words with the highest co-occurrence rate, the word selected is regarded as the correct meaning of the ambiguous word.

2.2 Method based on Semantic Class Dictionary

The arrangement of the genre dictionary is much different from the traditional dictionary. It is a dictionary that sorts and sorts words according to certain rules and interprets and discriminates similar words. store words that are similar to each other in the same directory. This makes it easy and quick to find synonyms. The unique hierarchy of semantic classification dictionaries makes it easier to provide semantic relationships between words. WordNet is a well-known dictionary of synonyms in English. Hownet and Harbin Institute of Technology's "Synonyms" [7] are both well-known Chinese-language dictionaries. When performing disambiguation tasks, Divide the dictionary into multiple levels according to the meaning of the words, then map the ambiguous words into the dictionary and find the word meaning classification that meets the conditions to complete the disambiguation of the words.

3. Supervised Methods

Supervised learning is a machine learning task that infers a function from tagged training data. The supervised learning algorithm analyzes the training data that has been given and produces an inferred function. Supervised Methods uses machine learning methods to perform word sense disambiguation using manually labeled training corpora. It performs well in the current disambiguation performance. This method is more discriminating against ambiguous words than Unsupervised Methods. The main reason is that this method can extract the semantic features well, and then you can classify the polysemous words. However, the size of the corpus will directly affect the performance of the classification. Building a high-quality corpus requires a lot of manpower, which makes it difficult for the method to make a big breakthrough in the corpus size. Furthermore, it is impossible to expand the scope of application of supervised word sense disambiguation. Furthermore, it is impossible to expand the application range of Supervised Methods. The following is a description of commonly used supervised machine learning models.

3.1 Bayesian Model

The Bayesian Model [8] is a typical probabilistic classification machine learning model. The Bayesian Model is a typical probabilistic classification machine learning model, which is a model that predicts the maximum probability of an unknown event in the current situation by the probability of a known event under given conditions [9]. In the disambiguation task, the conditional probability of the ambiguity word meaning with class mark feature is calculated, and then the appropriate word meaning classification is selected by the maximum conditional probability of the feature set tested. The word meaning with high probability is the correct word meaning of the ambiguity word.

The process of solving the disambiguation problem with the Bayesian model is shown in formula 1:

$$P(C_i|\text{Context}) = \frac{P(Context|S_i)P(C_i)}{\sum_{j=1}^{n} P(Context|C_j)P(C_j)} \quad (1)$$

3.2 Maximum Entropy Model

Maximum Entropy (ME) was proposed and established by E.T. Jaynes [10] based on the theory of information entropy. Its principle is that when it is necessary to predict the probability of occurrence of a random event, the prediction should satisfy all the pre-set conditions, but when it comes to uncertainty, it needs to retain various possibilities to solve. In this case, the probability distribution is the most uniform, and the probability of the predicted risk is minimal. The maximum entropy model has significant effects on multiple tasks in the field of natural language processing, such as: part-of-speech tagging, semantic role tagging, phrase recognition, etc [11]. P (C_{word}) is the probability distribution of the discrete random variable C_{word}. The process of entropy is shown in Equation 2:

$$H(P) = -\sum_{C_{word}} P(C_{word}) \log P(C_{word}) \quad (2)$$

The disambiguation process of the maximum entropy classifier is as follows: firstly, the sentences of training corpus and test corpus are processed by word segmentation and part-of-speech tagging, the lexical units adjacent to the ambiguous vocabulary are extracted, and the morphological and part-of-speech features of each lexical unit are extracted as disambiguation features. These characteristics are used to train the maximum entropy classifier, and then the word sense disambiguation is performed on the test corpus to obtain the classification result after disambiguation.

3.3 Conditional Random Field Model

Conditional Random Fields (CRF) [12] is a typical discriminant probability model proposed by Lafferty et al. in 2001. This model focuses on solving the problem of serialization annotation and is now commonly used in the study of part-of-speech tagging and named entity recognition [13]. CRF is an undirected graph model. The conditional random field is the conditional probability distribution model P(Y|X), which represents the Markov random field of another set of output random variables Y given a set of input random variables X. The vertices in the figure represent random variables, and the lines between the vertices represent the dependencies between random variables. In the conditional random field, the distribution of the random variable Y is the conditional probability, and the given observation is the random variable X. In principle, the layout of the conditional random field graph model can be arbitrarily given, as shown in Figure 1:

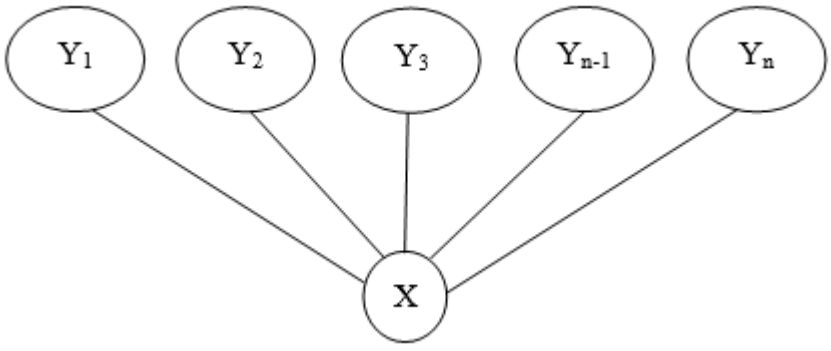

Figure 1. CRF chain structure diagram

X= {x_1, x_2, x_3…x_n} is an already given observation sequence, the probability form of the output label sequence is Y:

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_i \sum_k \lambda_k f_k(y_{i-1}, y_i, x)\right) \quad (3)$$

Z(X) is a naturalization factor; $f_k(y_{i-1}, y_i, x)$ is a feature function; Parameter λ_k represents the eigenvalue weight.

3.4 Support Vector Machine Model

Support Vector Machine (SVM) [14] is a machine learning algorithm developed on the basis of statistical learning theory. Compared with traditional algorithms, SVM has a solid mathematical foundation and can overcome the problems of dimensional disaster and over-fitting [15]. SVM is often used to solve classification problems. This method has been widely used in word sense disambiguation research in recent years. The basic principle of SVM is to find a hyperplane using the principle of interval maximization to segment the sample data and finally trans-form it into a convex quadratic programming problem. SVM can minimize structural risk and improve generalization ability, that is, less error is obtained from a limited training set sample. Due to this characteristic of SVM, the obtained local optimal solution must also be the global optimal solution. Classification sample: $(x_1, y_1), \ldots, (x_n, y_n)$ ($x_i \in R^n$, $y_i \in \{-1, +1\}$), The SVM decision function is:

$$g(x) = sign\left[\sum_{i=1}^{n} a_i y_i K(x_i, x) + b\right] \tag{4}$$

$K(x_i, x)$ is a kernel function, $K(x_i, x) = x_i \cdot x$.

3.5 Decision Tree Model

The Decision Tree [16] is a common machine learning model whose main job is to classify data [17]. Decision trees are often generated from the top down. The same problem is encountered at each node. Different answers to the questions on each node will result in different branches, so that the data is divided into multiple small subsets, until each subset contains only the same type of data, thus Get a decision tree.

Establishing a decision tree requires continuous segmentation of the data. Each segmentation corresponds to a problem and also corresponds to a node. Each segmentation is performed according to the principle of maximum difference. The generated decision tree can classify an unknown instance based on the value of the attribute. This algorithm is a form in which training data is represented as "attribute-conclusion". The non-leaf nodes of the decision tree are composed of attributes or attribute sets, and the leaf nodes are the categories of the categories. The instability of the decision tree makes it not widely used in word sense disambiguation. Minor changes to the pruning strategy and parameter settings will greatly affect the experimental results.

How to choose decision attributes is a very important problem in decision tree algorithm, the most famous of these solutions is the ID3 algorithm. The algorithm introduces information entropy in the judgment of decision attributes. This algorithm takes the rate of decline of information entropy as the criterion for selecting decision attributes. The attribute with the highest information gain is the decision attribute, and the technique of incremental learning is also introduced to improve the memory shortage caused by the excessive amount of training data. However, it cannot handle continuous attributes and the constructed decision tree over-fitting the training data. Therefore, the C4.5 algorithm appears. The algorithm selects the attribute with the highest information gain rate of C4.5 as the decision attribute and processes the continuous attribute with the information gain rate. In addition, the problem of overfitting is overcome by pruning.

4. Unsupervised Methods

Unlike Supervised Methods, Unsupervised Methods does not use a knowledge base such as an external dictionary for disambiguation. The most representative is the Lesk [18] algorithm. The algorithm is a dictionary-based word meaning elimination method, this algorithm considers that the meaning interpretation of a polysemous word has certain similarities with the sentence in which the polysemy is located, the main work of the Lesk algorithm is to measure this similarity. First, extract the phrase containing the ambiguous word in the sentence to be processed, and query each meaning

and interpretation information of the ambiguous word in the dictionary. Second, match each meaning of the ambiguous word with each meaning of the word in the context, and calculate the frequency of common words between the words in the phrase. Finally, the most frequently used meaning term is selected as the correct meaning of the ambiguous word.

Instead, it uses a corpus that has not been manually labeled to complete the disambiguation task. Unsupervised Methods mainly uses the corpus as a knowledge source, and uses clustering algorithms to cluster all the contextual word sets in which ambiguous words appear. This method usually does not label the meaning of the word, but can distinguish different meanings. The number of clusters is the semantic number of words. When a clustering result is obtained, the semantics are assigned to each cluster using a word meaning labeling algorithm. By using the semantic resources such as the semantic dictionary, the disambiguation work of the ambiguous vocabulary can be completed by aligning the sample examples of semantic recognition with the semantic resources.

Popescu [19] pointed out that a powerful clustering technique is enough to make up for the problems caused by insufficient external knowledge. Gaura [20] uses semantics to cluster and compare the similarity between the context of the ambiguous words and the predefined categories. The high similarity is used as the correct meaning of the ambiguous words. After SemEval proposed Cross-lingual Word Sense Disambiguation (CLWSD) [21], scholars tried to use bilingual or multilingual parallel corpus for disambiguation. The linguistic knowledge provided by these parallel corpora is helpful for semantic disambiguation, especially the linguistic knowledge that bilingual parallel corpora can provide. Kaji proposes an unsupervised disambiguation method based on a comparable bilingual corpus. Among them, SemEval-2013's Task3 provides parallel corpus of French, Dutch, German, Italian and Spanish as the English word sense disambiguation, Scholars use this knowledge to make word sense disambiguation. Lefever and Hoste [22] use five parallel corpora to construct a knowledge base of words and use this knowledge base for disambiguation; Gompel and Bosch [23] use the K-nearest neighbor classifier to map the local and global features of words to a parallel corpus, that is, the cross-language meaning of these words, using these cross-language meanings to eliminate; Also in this data set, Apidianaki[24] clusters all the corpora including five parallel corpora and uses the results of clustering to disambiguate.

5. Evaluation Index

The evaluation results of the experimental results of word sense disambiguation are mainly: separate the WSD task from the word, and do not consider the influence of the context of the ambiguous word, and evaluate the performance of the word sense disambiguation effect separately; the famous international word sense disambiguation contest Senseval uses This is the method of evaluation. The word sense disambiguation evaluation index uses the accuracy, recall rate, coverage rate and F1 value calculation methods commonly used in text classifiers to measure the performance of the system.

Precision indicates the accuracy of the word sense disambiguation task; Coverage indicates the coverage of the ambiguous words in the word sense disambiguation task; Recall indicates the ratio between the word mark disambiguation task and the N mark; F-1 refers to the inclusion of P And the combined value of R. N indicates the number of ambiguous words that have been marked, NT is the number of ambiguous words previously marked, and NR is the number of ambiguous words obtained in the experiment.

$$Precision = \frac{NR}{NT} \times 100\% \tag{5}$$

$$Coverage = \frac{NT}{N} \times 100\% \tag{6}$$

$$Recall = \frac{NR}{N} \times 100\% \tag{7}$$

$$F_1 = \frac{2PR}{P+R} \times 100\% \tag{8}$$

6. Summary

Among these three methods of word sense disambiguation, the accuracy of the disambiguation results of Knowledge-based Methods is higher than that of the other two disambiguation methods, but the disambiguation result is affected by the sparse problem of the data, which needs a huge and perfect Knowledge base; Supervised Methods has achieved good results in the word sense disambiguation problem, but in order to overcome the data sparse problem, this kind of method must have a larger standard corpus; Unsupervised Methods does not require a knowledge base and a manually annotated corpus. It can realize the training and learning of large-scale real corpus across domains, which can effectively overcome the data sparse problem, but its disambiguation result is the least desirable among the three types of disambiguation methods. The disadvantage of Unsupervised Methods is that the selection of disambiguation initial knowledge is often subjective and the accuracy is not stable enough. However, due to the scalability of Unsupervised Methods, it has a broader development space than Supervised Methods.

Acknowledgements

This work is partially supported by the National Natural Science Foundation of China (61363053).

References

[1]. NAVIGLI R, VELARDI P. Structural semantic interconnections: a knowledge-based approach to word sense disambiguation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005,27(7):1075-1086.

[2]. JIANGSHENG Y. Word sense disambiguation [J]. Computer Speech & Language, 2012, 36 (6);2355-2356

[3]. AGIREE E, EDMONNDS P. Word sense disambiguation [J]. Algorithm and Application, 2007 (10) :1-28.

[4]. Christiane Fellbaum. WordNet: An Electronic Lexical Database. MIT Press. 1998.

[5]. Zhendong Dong. HowNet [EB/OL]. http://www.Keenag.com, 2013.

[6]. SUBARANI M D. Concept based information retrieval from text documents[J]. Dept of Computer Sciences, 2012,2(4):38-48.

[7]. Jiazhen Mei, Yiming Yan. Synonym Word Forest [M]. Shanghai: Shanghai Dictionary Press, 1993: 106-108.

[8]. G. Escudero, L. Marquez, et al. Naive Bayes and Examplar Based Approaches to Word Sense Disambiguation Revisited. In Proceedings of the 14th Europear Conference on Artificial Intelligence, ECAI. 2000.

[9]. RASEKHAH, SADREDDINIMH, FAKHRAHMADSM. Word Sense Disambiguation Based on Lexical and Semantic Features Using Naive Bayes Classifier[J]. Journal of Computing and Security,2014,2(1):123-132.

[10]. Edwin Thompson Jaynes. Information Theory and statistical mechanics[J]. Physical Review, 1957, 106(4):620-630.

[11]. Adam L. Berger, Stephen A.Della Pietra, Vincent J.Della Pietra. A maximum entropy approach to natural language processing[J]. Computational Linguistics, 1996, 22(1):39-71.

Gaurav S Tomar, Manmeet Singh, Shishir Rai. Probabilistic latent semantic analysis for unsupervised word sense disambiguation[J]. International Journal of Computer Science Issues, 2013: 5(2),127-133.

[12]. Chengxiang Zhai, John Lafferty. A study of smoothing methods for language models applied to Ad Hoc information retrieval[J]. Acm Transactions on Information Systems, 2001,22(2):334--342.

[13]. John Lafferty, Andrew Mcallurn, Fernando Perieira. Conditioanl random field: Probabilistic models for segmenting and labeling sequence data[C]. In Proceeding of International Conference on Machine Learning, 2001:282-289.

[14]. Hideki Isozaki, Hideto Kazawa. Efficient Support Vector Classifiers for Named Entity Recognition. In proceedings of COLING-2002, Taipei, 2002: 390~396.

[15]. James Mayfield, Paul McNamee, and Christine Piatko. Named Entity Recognition Using Hundreds of Thousands of Features. In Walter Daelemans and Miles Osborne, editors,Proceedings of CoNLL-2003, Edmonton, Canada, 2003: 184~187.

[16]. Disambiguating Senseval Lexical Samples. In the Proceedings of the Workshop on Word Sense Disambiguation: Recent Successes and Future Directions. Philadelphia. July 11, 2002.

[17]. J. R. Quinlan. Induction of Decision Trees. Machine Learning. 1986, 1(1): 81~106.

[18]. LESK M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone[C]. Proceedings of the 5th Annual International Conference on Systems Documentation,1986:24-26.

[19]. POPESCU M, HRISTEA F. State of the Art Versus Classical Clustering for Unsupervised Word Sense Disambiguation[J]. Artificial Intelligence Review, 2011, 35(3): 241-264.

[20]. Gaurav S Tomar, Manmeet Singh, Shishir Rai. Probabilistic latent semantic analysis for unsupervised word sense disambiguation[J]. International Journal of Computer Science Issues, 2013: 5(2),127-133.

[21]. Tristan Miller, Nicolai Erbs, Hans-Peter Zorn, et al. DKPro WSD: A generalized UIMA-based framework for word sense disambiguation[C]. In Proceeding of ACL'13, 2013: 37-42.

[22]. Els Lefever, Veronique Hoste. Semeval-2010 task 3: Cross-lingua word sense disambiguation [C]. In Proceeding of ACL'10,2010:15-20.

[23]. Maarten van Gompel, Antal van den Bosch. WSD2: Parameter optimisation for memory-based cross-lingual word-sense disambiguation[C]. In Proceeding of ACL'13, 2013: 183-187.

[24]. Marianna Apidianaki. Data-driven synset induction and disambiguation for wordnet development [J]. Language Resources & Evaluation, 2009: 77-85.