International Journal of Information & Computation Technology.
ISSN 0974-2239 Volume 3, Number 1 (2013), pp. 25-31

http://www. irphouse.com

Statistical Approaches of Ambiguity Resolution in Natural Language Processing

***[1]Pankaj V. Nimbalkar and [2]Dr. P.K. Butey**

Assistant Professor,Dr. Ambedkar college,Nagpur-10
Email:pnimbalkar20@rediffmail.com
Associate Professor, Kamla Nehru Mahavidyalaya, Nagpur-27
Email: buteypradeep@yahoo.co.in

Abstract

"In this paper, we explores key idea of probability theory and how they might apply to natural language applications, some techniques for estimating probability from corpora and develops techniques for part of speech tagging. The goal of unsupervised learning is to group data into clusters. The main statistical techniques are mixture models and the expectation maximization (EM) algorithm"

Keyword : POS tagging, Hidden Markov Model, Naive Bayes

Introduction

The statistical approach to natural language processing (NLP) has become more and more important in recent years. This paper gives an overview of some fundamental statistical techniques that have been widely used in different NLP tasks. Methods for statistical NLP mainly come from machine learning, which is a scientific discipline concerned with learning from data[1]. That is, to extract information, discover patterns, predict missing information based on observed information, or more generally construct probabilistic models of the data. Machine learning techniques covered in this paper can be divided into two types: supervised and unsupervised [2].

Supervised learning is mainly concerned with predicting missing information based on observed information. For example, predicting part of speech (POS) based on sentences. It employs statistical methods to construct a prediction rule from labeled training data. Supervised learning algorithms discussed in this paper include naive Bayes, support vector machines (SVMs), and logistic regression[3]. The goal of unsupervised learning is to group data into clusters. The main statistical techniques are mixture models and the expectation maximization (EM) algorithm. This paper will

also cover methods used in sequence analysis, such as hidden Markov model (HMM), conditional random field (CRF), and the Viterbi decoding algorithm[4].

A statistical model is probability distribution over all possible word sequences.
n-gram model:-The goal of statistical language model is to estimate the probability of a sentence. This is achieved by decomposing sentence probability into a product of conditional probabilities using chain rule

p(s)=p(w1,w2,……..,wn)
=p(w1)p(w2/w1)p(w3/w1w2)…p(wn/w1w2…….wn-1)
=Πp(wi/hi)

Where '*hi*' is history of word

In order to calculate sentence probability, we need to calculate the probability of word, given the sequence of words preceding it. An n-gram model simplifies the task by approximating the probability of word given all the previous words by the conditional probability given previous n-1 words only.

An n-gram model calculates probability by modeling language as Markov model of order n-1 words i.e. by looking n-1 words only .A model that limits the history to the previous one word only is termed bi-gram model. A model that conditions the probability of a word to the previous two words is called trigram model.

A special word <s> is introduced to mark the beginning of the sentence in bi-gram estimation .similarly n-gram model parameters should be estimated using maximum likelyhood estimation (MLE) technique. The sum of all n-grams that share first n-1 words is equal to the count of the common prefix.

Key ideas in probability Theory:- Probability can be defined in terms of random variable, which may range over predefined set of values .while random variables may range over infinite sets and continuous values .Bayes's rule is very important in conditional probability conditional probability is given by

PROB(A/B)=PROB(B/A)*PROB(A)/PROB(B)

The statistical framework
The statistical framework attempts to answer the question: "Which of all of the possible Combinations of transfer mappings whose source component is covered by a portion of the input dependency graph will yield the best target language dependency graph?" The rest of the system components (i.e. the alignment and partitioning of the training data and the analysis and generation components) did not have to be modified for the new system.

The traditional noisy-channel SMT model attempts to find the highest-probability translation for a sentence

$T = \arg\max(P(T \mid S))$, (1)

where S is the source language sentence and T the target language sentence. By Bayes' rule,

$T = \arg\max(P(T \mid S)) = \arg\max(P(S \mid T)P(T))$. (2)

A target language model trained on monolingual target language data is used to compute an estimate of P(T), and channel models of varying complexity are built to compute and estimate P(S|T).In the system, individual candidate mappings and combinations of mappings are scored using a linearly interpolated combination of scores from several heterogeneous information sources. This "kitchen sink" approach to SMT is superficially similar to the system described in (Och and Ney, 2002), except that it works with dependency graph mappings instead of with surface strings.

Estimating probability

One method of estimation is to use the ratios from the corpus as the probability to predict the interpretation of the new sentence. For Ex:-We have an acceptable margin of error 0.4 and 0.6 for estimating the probability of a fair coin coming up heads, the chance of obtaining a reliable estimate for the probability with five flips is 89% by maximum likelihood estimator formula. But unfortunately there are vast numbers of estimate needed for Natural language application. The worst case occurs if a low frequency word does not occur at all in one of its possible categories.

Part of speech tagging

From a linguistic point of view, the linguists mostly agree that there are three major (primary) parts of speech: noun, verb, and adjective (Pustet, 2003).Part-of –speech tagging involves selecting the most likely of syntactic categories for the words in a sentence [5].

Part-of-Speech Tagging Approaches

Rule-Based Approaches

The earliest POS tagging systems are rule-based systems, in which a set of rules is manually constructed and then applied to a given text. Probably the first rule-based tagging system is given by Klein and Simpson (1963), which is based on a large set of handcrafted rules and a small lexicon to handle the exceptions. The initial tagging of the Brown corpus was also performed using a rule-based system, TAGGIT (Manning and Schütze, 2002). The lexicon of the system was used to constrain the possible tags of a word to those that exist in the lexicon. The rules were then used to tag the words for which the left and right context words were unambiguous. The main drawbacks of these early systems are the laborious work of manually coding the rules and the requirement of linguistic background [6].

Markov Model Approaches

The rule-based methods used for the POS tagging problem began to be replaced by stochastic models in the early 1990s. The major drawback of the oldest rule-based systems was the need to manually compile the rules, a process that requires linguistic background. Moreover, these systems are not robust in the sense that they must be partially or completely redesigned when a change in the domain or in the language

occurs. Later on a new paradigm, statistical natural language processing, has emerged and offered solutions to these problems. As the field became more mature, researchers began to abandon the classical strategies and developed new statistical models. The statistical POS tagging enables us to use statistical methods is the availability of a rich repertoire of data sources: lexicons (may include frequency data and other statistical data), large corpora (preferably annotated),bilingual parallel corpora, and so on. By using such resources, we can learn the usage patterns of the tag sequences and make use of this information to tag new sentences. We devote the rest of this section and the next section to statistical POS tagging models [7].

Maximum Entropy Approaches

The HMM framework has two important limitations for classification tasks such as POS tagging: strong independence assumptions and poor use of contextual information. For HMM POS tagging, we usually assume that the tag of a word does not depend on previous and next words, or a word in the context does not supply any information about the tag of the target word. Furthermore, the context is usually limited to the previous one or two words. Although there exist some attempts to overcome these limitations, Maximum entropy (ME) models provide us more flexibility in dealing with the context and are used as an alternative to HMMs in the domain of POS tagging. The use of the context is in fact similar to that in the TBL framework. A set of feature templates (in analogy to rule templates in TBL) is predefined and the system learns the discriminating features by instantiating the feature templates using the training corpus. The flexibility comes from the ability to include any template that we think useful—may be simple (target tag ti depends on ti−1) or complex (ti depends on ti−1 and/or ti−2 and/or wi+1). The features need not be independent of each other and the model exploits this advantage by using overlapping and interdependent features[8]

Taggers Based on ME Models

The flexibility of the feature set in the ME model has been exploited in several ways by researchers.Toutanova and Manning (2000) concentrate on the problematic cases for both unknown/rare words and known words. Two new feature templates are added to handle the unknown and rare words:

- A feature activated when all the letters of a word are uppercase
- A feature activated when a word that is not at the beginning of the sentence contains an uppercase letter, the distribution of words in which only the initial letter is capitalized is different from the distribution of words whose all letters are capitalized. Thus, such features need not be useful in other corpora

Other Statistical and Machine Learning Approaches

There are a wide variety of learning paradigms in the machine learning literature (Alpaydın, 2004).However, the learning approaches other than the HMMs have not been used so widely for the POS tagging problem. This is probably due to the suitability of the HMM formalism to this problem and the high success rates obtained with HMMs. Nevertheless, all well-known learning paradigms have been applied to

POS tagging in some degree. In this section, we list these approaches and cite a few typical studies that show how the tagging problem can be adapted to the underlying framework. The interested reader should refer to this chapter's section in the companion wiki for further details.

HMM-Based Taggers

A comprehensive analysis of the effect of using MMs for POS tagging was given in an early work by Merialdo (1994). In this work, a second-order model is used in both a supervised and an unsupervised manner. An interesting point of this study is the comparison of two different schemes in finding the optimal tag sequence of a given (test) sentence. The first one is the classical Viterbi approach as we have explained before, called "sentence level tagging" in Merialdo (1994). An alternative is "word level tagging" which, instead of maximizing over the possible tag sequences for the sentence, maximizes over the possible tags for each word:

Merialdo (1994) uses a form of interpolation where trigram distributions are interpolated with uniform distributions. A work that concentrates on smoothing techniques in detail is given in Sündermann and Ney (2003). It employs linear interpolation and proposes a new method for learning λi's that is based on the concept of training data coverage (number of distinct n-grams in the training set). It argues that using a large model order (e.g., five) accompanied with a good smoothing technique has a positive effect on the accuracy of the tagger. Another example of a sophisticated smoothing technique is given in Wang and Schuurmans (2005). The idea is exploiting the similarity between the words and putting similar words into the same cluster. Similarity is defined in terms of the left and right contexts. Then, the parameter probabilities are estimated by averaging, for a word w, over probabilities of 50 most similar words of w. the distribution of the unknown words is similar to that of the less probable words (words occurring less than a threshold t, e.g., t = 10).Therefore, the parameters for the unknown words can be estimated from the distributions of less probable words. Several models were tested, particularly first- and second-order HMMs were compared with a simpler model, named Markovian language model (MLM), in which the lexical probabilities P(W|T) are ignored. All the experiments were repeated on seven European languages. The study arrives at the conclusion that HMM reduces the error almost to half in comparison to the same order MLM.

Combining Taggers

POS tagging problem was approached using different machine learning techniques and 96%–97% accuracy seems a performance barrier for almost all of them. It was observed that, although different taggers have similar performances, they usually produce different errors (Brill and Wu, 1998; Halteren et al., 2001). Based on this encouraging observation, we can benefit from using more than one tagger in such a way that each individual tagger deals with the cases where it is the best.

One way of combining taggers is using the output of one of the systems as input to the next system. An early application of this idea is given in Tapanainen and

Voutilainen (1994), where a rule-based system first reduces the ambiguities in the initial tags of the words as much as possible and then an HMM-based tagger arrives at the final decision. The intuition behind this idea is that rules can resolve only some of the ambiguities but with a very high correctness and the stochastic tagger resolves all ambiguities but with a lower accuracy. The method proposed in Clark et al. (2003) is somewhat different and it investigates the effect of co-training, where two taggers are iteratively retrained on each other's output. The taggers should be sufficiently different (e.g., based on different models) for co-training to be effective. This approach is suitable in cases when there is a small amount of annotated corpora. Beginning from a seed set (annotated sentences), both of the taggers (T1 and T2) are trained initially. Then the taggers are used to tag a set of unannotated sentences. The output of T1 is added to the seed set and used to retrain T2; likewise, the output of T2 is added to the seed set to retrain T1. The process is repeated using a new set of unannotated sentences at each iteration. The second way in combining taggers is letting each tagger to tag the same data and selecting one of the outputs according to a voting strategy.

Conclusion

In addressing the ambiguity resolution statistically based methods show great promise in Natural Language processing[8].Viterbi algorithm using bigram and trigram probability models can attain accuracy rates of over 95 percent after experimental result. POS tagging problem t techniques divide into two broad categories: rule-based methods and statistical methods. The HMM framework is the most widely used statistical approach for the POS tagging problem.

References

[1] Banerjee, S. and Lavie, A. (2005). Meteor: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of Workshop on Intrinsic and Intrinsic Evaluation Measures for MT and Summarization at the 43rd Annual Meeting of the Association of Computational Linguistics (ACL-2005), Ann Arbor, MI, pp. 65–72.

[2] Block, H. U. (2000). Example-based incremental synchronous interpretation. In Wahlster, W., editor,Verbmobil: Foundations of Speech-to-Speech Translation, pages 411–417. Springer-Verlag, Berlin,Germany.

[3] Altunyurt, L., Z. Orhan, and T. Güngör. 2007. Towards combining rule-based and statistical part of speech tagging in agglutinative languages. Computer Engineering 1(1):66–69.

[4] Black, E., F. Jelinek, J. Lafferty, R. Mercer, and S. Roukos. 1992. Decision tree models applied to the labeling of text with parts-of-speech. In HLT, pp. 117–121, New York. ACL.

[5] Brill, E. and J. Wu. 1998. Classifier combination for improved lexical disambiguation. In COLING-ACL,pp. 191–195, Montreal, QC. ACL/Morgan

Kaufmann. Cao, H., T. Zhao, S. Li, J. Sun, and C. Zhang. 2005. Chinese pos tagging based on bilexical co-occurrences. In ICMLC, pp. 3766–3769, Guangzhou, China. IEEE.

[6] Darroch, J.N. and D. Ratcliff. 1972. Generalized iterative scaling for log-linear models. The Annals of Mathematical Statistics 43(5):1470–1480.

[7] Eineborg, M. and N. Lindberg. 2000. ILP in part-of-speech tagging—An overview. In LLL, eds. J. Cussens and S. Džeroski, pp. 157–169, Lisbon, Portugal. Springer.

[8] Finch, A. and E. Sumita. 2007. Phrase-based part-of-speech tagging. In NLP-KE, pp. 215–220, Beijing, China. IEEE.

[9] Grzymala-Busse, J.W. and L.J. Old. 1997. A machine learning experiment to determine part of speech from word-endings. In ISMIS, pp. 497–506, Charlotte, NC. Springer.

[10] Habash, N. and O. Rambow. 2005. Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In ACL, pp. 573–580, Ann Arbor, MI. ACL.

[11] Květon, P. and K. Oliva. 2002b. (Semi-)Automatic detection of errors in pos-tagged corpora. In COLING,pp. 1–7, Taipei, Taiwan. ACL.

[12] Lafferty, J., A. McCallum, and F. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In ICML, eds. C.E. Brodley and A.P. Danyluk, pp. 282–289, Williamstown, MA. Morgan Kaufmann.

[13] Murata, M., Q. Ma, and H. Isahara. 2002. Comparison of three machine-learning methods for Thai part-of-speech tagging. ACMTransactions on Asian Language Information Processing 1(2):145–158.

[14] Nagata, M. 1999. A part of speech estimation method for Japanese unknown words using a statistical model of morphology and context. In ACL, pp. 277–284, College Park, MD

[15] Oliva, K., M. Hnátková, V. Petkevič, and P. Květon. 2000. The linguistic basis of a rule-based tagger for Czech. In TSD, eds. P. Sojka, I. Kopeček, and K. Pala, pp. 3–8, Brno, Czech Republic. Springer.

[16] Padró, L. 1996. Pos tagging using relaxation labelling. In COLING, pp. 877–882, Copenhagen, Denmark.

[17] Roth, D. and D. Zelenko. 1998. Part of speech tagging using a network of linear separators. In COLINGACL,pp. 1136–1142, Montreal, QC. ACL/Morgan Kaufmann.

[18] Sak, H., T. Güngör, and M. Saraçlar. 2007. Morphological disambiguation of Turkish text with perceptron algorithm. In CICLing, ed. A. Gelbukh, pp. 107–118, Mexico. Springer.