A survey of machine learning approaches to analysis of large corpora

Xunlei Rose Hu <rosehu@comp.leeds.ac.uk>
and Eric Atwell <eric@comp.leeds.ac.uk>

School of Computing, University of Leeds, U.K. LS2 9JT

Abstract

Corpus-based Machine Learning of linguistic annotations has been a key topic for all areas of Natural Language Processing. This paper presents a survey, along three dimensions of classification. First we outline different linguistic level of analysis: Tokenisation, Part-of-Speech tagging, Parsing, Semantic analysis and Discourse annotation. Secondly, we introduce alternative approaches to Machine Learning applicable to linguistic annotation of corpora: N-gram and Markov models, Neural Networks, Transformation-Based Learning, Decision Tree learning, and Vector-based classification. Thirdly, we examine a range of Machine Learning systems for the most challenging level of linguistic annotation, discourse analysis; these illustrate the various Machine Learning approaches. Our overall aim is to provide an ontology or framework for further development of our research.

Key Words: Machine Learning, corpus, annotation, tagging, linguistic analysis, dialogue

1 Introduction

In the research area of Natural Language Processing, there have been significant advances in Machine Learning approaches for automatic analysis of corpora, covering a range of linguistic levels including: Tokenisation, Part-of-Speech tagging, Partial parsing, Semantic analysis and Discourse annotation.

Our goal is to develop a generic framework, a general-purpose toolkit which could be used to apply across all the linguistic levels (Atwell 1996). This approach aims to reduce or remove the need for human expert intervention to validate and enrich the output via post-editing. Then there is no longer a need to produce and store annotation files. Instead, we store only the raw corpus text, and the linguistic annotation of different layers can be generated (and regenerated) dynamically on demand. This is particularly appropriate for Surface Processing of Large Corpora (SproLaC); for "monitor corpora" which are continuously updated to reflect current language use; and for corpus-based research which uses continuously-changing online resources on the World Wide Web as a corpus, such as WebCorp.

As human beings, we understand natural language with such ease that we hardly pay attention to what "understanding" really involves. In recent years, the area of "intension" has become an increasingly important area of study. A manifestation of this is the trend toward the study of pragmatics and discourse analysis. In this survey, we are particularly interested in machine learning approaches to this "top level" linguistic analysis task. We begin by outlining the range of linguistic analysis levels amenable to Machine Learning approaches; then we describe the general types of machine learning algorithms which apply in language learning systems; and finally we look at some systems to deal with the more challenging problems of discourse analysis

2 Levels of linguistic analysis

2.1 Tokenisation

Tokenisation is the process of breaking up the sequence of characters in a text by locating the word boundaries, the points where one word ends and another begins (Palmer 2000). This was not seen to be a serious problem for researchers working on English and similar languages, where word boundaries are generally coincident with space characters; but it is a more challenging task for Chinese and some other languages: in Chinese, a word may be a single character, or a series of two or more characters, and there may be no spaces to separate words.

Tokenisation reseachers started by creating and maintaining hand-built algorithms since most of them processed small texts in a single language. However, the recent explosion in availability of large unrestricted corpora has forced a move toward developing corpus-based Machine Learning algorithms, which are more robust and do not depend on the well-formedness of texts being processed.

A common approach to tokenisation of Chinese is to start by considering each character a distinct word, and then to use a variation of the maximum matching algorithm, also called “greedy” algorithm.

2.2 Part-of-Speech tagging

In PoS-tagging, each word must be assigned its correct Part-of-Speech, such as noun, verb, adjective or adverb; furthermore, most PoS-taggers also give additional grammatical features, such as singular/plural number, tense, gender. The number of tags used by different systems varies a lot. Some systems use fewer than 20 tags, while others use over 400.

Many systems for Part-of-Speech tagging have learnt statistical models from a training corpus. An early example was CLAWS, the Constituent Likelihood Automatic Word-tagging System (Leech at al 1983a,b, Atwell 1983), which learnt a “Constituent Likelihood” model from the tagged Brown Corpus; CL was basically a tag-bigram model augmented with longer n-grams for special-case idioms. Stochastic taggers can achieve good results, around 93-97% accuracy. But stochastic taggers require tables of lexical and PoS ngram statistics and do not explicitly represent intuitive rules. Implementing improvements to the tagger can be difficult: a number of more sophisticated variants on the original CL PoS-bigram model have been tried, but without major improvements in accuracy.

The most popular alternative Machine Learning approach to PoS-tagging is to use Transformation-Based Learning (Brill 1995) to learn local tag-combination constraint rules, which are used to eliminate candidate tags incompatible with immediate context, or to select a tag which is required in a specific context. The Brill Tranformation-Based tagger has been retrained with a range of tagged corpora to produce a range of PoS-taggers, eg see (Atwell et al 2000).

2.3 Parsing

The task of an automatic parser is to take a formal grammar and a sentence and apply the grammar to the sentence, to produce a parse-tree structure for the sentence. Parsing is a well established NLP task. Two contrasting starting-points reflect contrasting perspectives: one starts out with the words of the sentence, and builds the tree `bottom up'; the other starts with the S node, builds the tree `top-down'. Bottom-up parsing can be approached as a variation of tagging, adding higher-level tags on top of PoS-tags. For example, (Atwell 1983, Atwell et al 1984) represents a parse-tree as a sequence of labelled brackets, and a higher-level tag or “hypertag” is a bundle of closing and/or opening brackets between two PoS-tags; a new sentence must first be PoS-tagged by CLAWS and then the PoS-tags are used to predict hypertags. A top-down parser can be machine-leant from a hand-parsed corpus by extracting all subtrees of constituent + immediate daughters, and converting each of these into a context-free rule. For example, (Atwell 1985) learned a Prolog Definite Clause Grammar parser from and-parsed samples of the LOB Corpus. Improvements in computer speed and memory have allowed Machine Learning of larger patterns from a parsed corpus; for example, DOP: Data Oriented Parser (Bod 1993)

learns all possible substructures from a parsed corpus, and parses a new sentence by finding the optimal combination of subtrees to span the sentence.

2.4 Semantic Analysis

Semantic annotation is augmentation of data to facilitate automatic recognition of the underlying semantic content and structure. A common practice in this respect is labelling of documents with thesaurus classes for the sake of document classification and management. There is no universal agreement about which semantic features ought to be annotated (Demetriou and Atwell 2001) - in fact in the past much of the annotation was motivated by linguistic theories of, for instance, social interaction. Semantic annotation has been used in connection with machine-learning software trainable on annotated corpora for word-sense disambiguation, co-reference resolution, summarization, information extraction, measuring semantic similarity or difference between documents, and other tasks.

2.5 Discourse Analysis

Part-of-Speech tagging, parsing, and semantic analysis take place at the level of word and sentence: each sentence in a text can be analysed independently. Language in real use in spoken dialogues exhibits structure beyond sentence-boundaries. In a conversation, meanings can refer back to previous sentences; and a series of turns between two speakers generally has an overall structure to meet the goals of the participants. As with semantic categories, there is no universal agreement on discourse analysis categories or labels; but a growing range of dialogue transcript corpora have been hand-annotated with dialogue-act or speech-act tags designed for specific applications. These are interesting training sets for Machine Learning research, because the sequencing or pattern of dialogue acts is not straightforwardly predictable from past dialogue act(s) alone.

3 Machine Learning techniques for linguistic annotation of corpora

3.1 N-gram and Markov models

A Markov model of a sequence of states or symbols (eg words or Part-of-Seech tags) is used to estimate the probability or “likelihood” of a symol sequence. It can be used for disambiguation, eg for choosing the “likeliest” tag for an ambiguous word in a given context, by estimating the probability of every candidate sequence. A Markov model applies the simplifying assumption that the probability or “likelihood” of a long sequence or chain of symbols can be estimated in terms of its parts or n-grams.

Hidden Markov Models (HMMs) are a variant including 2 layers of states: a visible layer corresponding to input symbols (eg words) and a hidden layer learnt by the system, corresponding to broader categories (eg wod-classes).

As mentioned above, Markov or n-gram models have been widely used for Part-of-Speech tagging, following the successful use in tagging the LOB Corpus (Leech et al 1983).

3.2 Neural Networks

Neural networks have been widely explored in Artificial Intelligence (AI). NNs have been studied for many years in the hope of achieving human-like performance in many fields.

There are many rules used in the learning process of neural networks. The type of learning in a neural network is determined by the manner in which the parameters change. This can happen with or without a teacher; hence, the neural networks are divided into three groups: supervised learning, unsupervised learning and reinforcement learning.

A related model is the semantic network. Semantic networks typically have nodes that represent concepts and connections that represent semantically meaningful associations between these concepts. They are better characterized as associative network models than as neural/brain models. The activation rules that implement information retrieval in these associative networks, often referred to as spreading activation, typically produces an intersection search. Hence, they are also called "spreading activation" models (Doszkocs, Reggia and Lin, 1990).

3.3 Transformation-Based Learning

Brill (1995) developed a symbolic Machine Learning method called Transformation-Based Learning (TBL). Given a tagged training corpus, Transformation-Based Learning produces a sequence of rules that serves as a model of the training data. To derive the appropriate tags, each rule may be applied in order to each instance in an untagged corpus.

TBL relies heavily on a large annotated training corpus, and relies on reasonable default heuristics to get things started. It learns rules that are easily understandable and allows rules to be easily acquired for different domains or genres. As mentioned above, TBL has been widely used for Part-of-Speech tagging, eg (Atwell et al 2000).

There is a gap between an initial semantic network generated from input data, and a semantic one representing profound knowledge, from which a knowledge database can be constructed. By using transformation rules, the semantic analysis method is based on a pattern matching with a semantic network. A transformation rule description language allows users to manipulate their knowledge base and to define rules.

3.4 Decision Tree classification

A decision tree is constructed by recursively partitioning the training set, selecting, at each step, the feature that most reduce the uncertainty about the class in each partition, and using it as a split.

For example, (Cohen 1995) used the decision tree learner Ripper to induce a decision tree that was used to automatically label a new corpus with predicates, and used 5-fold cross validation to ensure results were stable.

3.5 Vector-based clustering

This approach uses co-occurrence statistics to construct vectors that represent word classes or meanings by virtue of their direction in multi-dimensional word-collocation space. For example, (Atwell 1983) annotated each word in a sample from the LOB Corpus with a vector of neighbouring word-types; words with similar vectors were clustered into word-classes.

A method for calculating semantic word vectors is to use random labelling of words in narrow context windows to calculate semantic context vectors for each word type in the text data. Incorporating linguistic information in the context vectors can enhance the results.

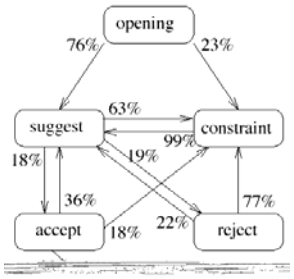
4 Discourse Analysis

Although Machine Learning approaches have achieved success in many areas of Natural Language Processing, researchers have only recently begun to investigate applying Machine Learning methods to discourse-level problems. To interpret an utterance's dialogue act is an important task in discourse understanding, which is a concise abstraction of the speaker's intention. This section surveys approaches of discourse annotation at the Dialogue Act (DA) level. Recognizing DAs is critical for discourse-level understanding and can also be useful for other applications, such as resolving ambiguity in speech recognition.

Markov models and Hidden Markov models are popular in pattern detection in discourse analysis. Almost every project discussed has investigated n-grams and/or HMMs. Neural networks and decision trees have been experimented with and found largely effective. Brill's transformation-based learning offers an alternative to HMMs.

4.1 Woszczyna and Waibel (1994): N-grams, Markov model

This is an early attempt at dialogue act tagging. There are six dialogue acts and their transition probabilities, shows in the figure below.



Woszczyna and Waibel also calculated "emission probabilities". This system was programmed to group the input vocabulary into clusters with similar contexts, and then worked with the clusters. The classes of words built by the system worked comparably to groups that the experimenters hand-built.

4.2 Reithinger, Engel, Kipp, and Klesen (1996): N-grams, HMM

The experiment was carried out for German and for English. It gave satisfactory results with n-grams and speaker information, performing the recognition with a hidden Markov chain.

This system used a hidden Markov chain to detect the patterns of dialogue. Reithinger et al. used a form of dialogue grammar called *word n-grams*, or strings of words *n* words long, as input. It was discovered that tri-grams are most effective in recognising DAs for the centres of utterances, and that bi-grams are most effective in recognising DAs for the beginnings and ends of utterances.

4.3 Mast et al. (1996): Decision tree, n-grams

This study was similar to (Reithinger et al 1996), but with an addition of a GARBAGE class for uninterpretable utterances, false starts, and so on.

Two methods were used:

- Semantic classification trees (SCTs). An SCT is a kind of decision tree that uses word patterns.
- Polygram language models (LMs), similar to n-grams.

4.4 Reithinger and Klesen (1997): n-gram, Bayesian network

Reithinger and Klesen used Verbmobil, and the data was transcribed and segmentated. This work heavily relied on word n-grams. It performed poorly on DA's that occurred infrequently in the corpus. An exception to this was SUGGEST, which appeared frequently in the corpus, but was not recognised as well because the wording was so different between instances. The team got better results for English than for German, probably because English word order is more fixed.

They also performed pattern detection using Bayesian networks. As with HMMs, DAs with fixed phrases approached 85-100% recognition, compared to 60-65% for others. Digressions from topic were recognised most poorly.

4.5 Samuel, Carberry, and Vijay-Shanker (1998): Transformation-Based Learning

Samuel, Carberry and Vijay-Shanker developed a Monte Carlo version of the TBL algorithm that coped with the large-scale search problem of selecting constraints from all possible combinations of a pre-selected set of conditions. Each condition consists of a feature and a distance. The feature specifies a characteristic of utterance that might be relevant for the Dialogue Act Tagging task, and the distance specifies the relative position of the utterance that the feature should be applied to.

In order to identify the phrases that will be useful for a particular domain, they uses a statistical approach to select relevant cue patterns from a training corpus and analysed the distribution of dialogue acts for utterances that include a given phrase. When using these cue patterns, the system's accuracy rose by 18%.

4.6 Wright (1998): n-gram, CART, neural networks

Wright was especially interested in how predicting Dialogue Acts (DAs) could help in speech recognition by constraining the possible range of candidate words. Her goal was to tag spontaneous speech using prosody. She used the DCIEM Map Task corpus, with Canadian speakers. According to the Dialogue Act model used, conversations are viewed as a series of 'games', and are analysed for the 'moves' made by players. These moves correspond to the twelve DA tags for the Map task corpus.

Wright also examined DA *n-grams*, and claimed that 4-grams have the most predictive power. Speaker information was also examined. When examining the strings of DAs, Wright looked at current speaker role (whether initiator or not); the move type of the *other* speaker's most recent move; and the role of the speaker of the immediately preceding move (which may be the same as 2). Wright compared three different kinds of pattern recognition: hidden Markov models, classification and regression trees (CARTs), and artificial neural networks (NNs). All methods performed comparably.

4.7 Taylor, King, Isard, and Wright (1998): combined n-grams and HMM

The features under study were as follows:

- Intonation: Intonation was a reliable predictor of DA in many cases.
- Syntactic features of words: Work in syntactic forms using wh-questions.
- DA n-grams: 4-grams were found to have the greatest predictive power.

This system combines the likelihood from all three of the above models and calculates the most likely DA. An HMM is used for pattern detection.

4.8 Fukada et al. (1998): bi-gram, HMM

The DA definition in this work is a combination of speaker information, speech act, concept, and argument.

The system was experimented with using two corpora; one in Japanese (collected by ATR) and one in English (collected by Carnegie Mellon University). Both corpora dealt with travel arrangement tasks.

Before processing, some words in the corpus were classified into syntactic groups. Other words ("uh", "erm") were dropped completely.

This work relied heavily on bi-grams. HMMs were used in pattern recognition. Results were positive; the DA tagger for Japanese tagged the correct DA 81% of the time, higher than any previous attempt. Accuracy for English, however, was 57%, well below the benchmark set by other similar projects.

4.9 Stolcke et al. (1998): HMM, decision trees

The tag set is a modified version of the DAMSL tagset. About 220 of the possible combinations appeared in the corpus. The researchers then grouped some of these combinations into a less fine-grained tagset, comprising 42 DAs in all.

The team used HMMs for pattern recognition. The features they focused on were DA n-grams and speaker information. The researchers also used prosody to help in the classification process. Decision trees were used for this, though the researchers note some possible advantages of using other methods such as Neural Networks. Interestingly, the team also included utterance length in the process, though utterance length is not strictly a prosodic feature.

5 Conclusion

This paper presents a survey of a range of approaches, classified on the dimension of different linguistic level of analysis: Tokenisation, Part-of-Speech tagging, Parsing, Semantic analysis and Discourse annotation. It provides a framework for the development of new components.

Our research plan is to develop a generic toolkit applied across all the linguistic levels. This survey has explored all the underlined algorithms in different linguistic levels. It provides an ontology or framework for further development of our research.

Some systems can be used and reused in different linguistic levels, such as systems based on Brill Transformation-Based Learning, Decision Tree, etc. In the future, we aim to integrate such systems and comparatively evaluate Machine Learning corpus analysis techniques.

6 References

Atwell, E, 1983. *Constituent-Likelihood Grammar.* ICAME Journal of the International Computer Archive of Modern English Vol.7

Atwell, E, Leech, G & Garside, R, 1984. *Analysis of the LOB Corpus: progress and prospects* in Aarts, J & Meijs, W (editors), Corpus Linguistics: Proceedings of the ICAME 4th International Conference on the Use of Computer Corpora in English Language Research, pp40-52, Rodopi.

Atwell, E, 1988. *Transforming a Parsed Corpus into a Corpus Parser* in Kyto, M, Ihalainen, O & Risanen, M (editors), Corpus Linguistics, Hard and Soft: Proceedings of the ICAME 8th International Conference on English Language Research on Computerised Corpora, pp61-70, Rodopi.

Atwell E. 1996. *Machine Learning from Corpus Resources for Speech And Handwriting Recognition* in Thomas J, and Short M (editors), *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*, pages 151-166, Longman, Harlow.

Atwell E, Demetriou G, Hughes J, Schiffrin A, Souter C, and Wilcock S. 2000. *A comparative evaluation of modern English corpus grammatical annotation schemes*. ICAME Journal, volume 24, pages 7-23, International Computer Archive of Modern and medieval English, Bergen.

Bod, R, 1993. *Using an annotated corpus as a stochastic grammar.* In Proceedings of EACL, the sixth conference of the European chapter of the Association for Computational Linguistics, pp.37-44.

Brill, E. 1995. *Transformation-based error-driven learning and natural language processing: a case study in Part-of-Speech tagging*. Computational Linguistics, volume 21(4), pages 543-566.

Cohen, P, 1995. *Empirical methods for Artificial Intelligence.* MIT Press, Cambridge MA.

Demetriou G and Atwell E. 2001. *A domain-independent semantic tagger for the study of meaning associations in English text*. In Harry Bunt, Ielka van der Sluis and Elias Thijsse (editors), Proceedings of the Fourth International Workshop on Computational Semantics (IWCS-4) pp.67-80. Tilburg, Netherlands.

Leech, G, Garside, R & Atwell, E, 1983 *Recent developments in the use of computer corpora in English language research*, in Transactions of the Philological Society, pp.23-40.

Leech, G, Garside, R & Atwell, E, 1983. *The Automatic Grammatical Tagging of the LOB Corpus* ICAME Journal of the International Computer Archive of Modern English Vol.7

Mast, M, Kompe, R, Harbeck, S, Keissling, A, Niemann, H, Noth, E, Schukat-Talamazzini, E, and Warnke, V, 1996. *Dialog act classification with the help of prosody*. In Proceedings of ICLSP-96, Philadelphia, volume 3, pp1732-1735.

Reithinger, N, Engel, R, Kipp, M, and Klesen, M 1996. *Predicting dialogue acts for a speech-to-speech translation system.* In Proceedings ICLSP-96, Philadelphia, volume 2, pp.654-657.

Reithinger, N, and Klesen, M, 1997. *Dialogue act classification using language models.* In Proceedings of EUROSPEECH-97, volume 4, pp.2235-2238.

Samuel, K, Carberry, S, and Vijay-Shanker, K, 1998. *Dialogue act tagging with transformation-based learning.* In Proceedings of COLING/ACL-98, Montreal, volume 2, pp.1150-1156.

Stolcke, A, Shriberg, E, Bates, R, Coccaro, N, Jurafsky, D, Martin, R, Meteer, M, Ries, K, Taylor, P, and Van Ess-Dykema, C. 1998. *Dialog Act Modeling for Conversational Speech*, in Chu-Carroll, J, and Green, N, (eds) Applying machine learning to discourse processing: AAAI Spring Symposium, pp.98-105.

Taylor, P, King, S, Isard, S, and Wright, H, 1998. *Intonation and dialog context as constraints for speech recognition.* Language and Speech, volume 41(3-4), pages 489-508.

Woszczyna, M and Waibel, A, 1994. *Inferring linguistic structure in spoken language.* In ICSLP-94, Yokohama, pp.1363-1366.