Chapter 14
Machine Learning in Natural Language Processing

Marina Sokolova
CHEO Research Institute, Canada

Stan Szpakowicz
University of Ottawa, Canada and Polish Academy of Sciences, Poland

ABSTRACT

This chapter presents applications of machine learning techniques to traditional problems in natural language processing, including part-of-speech tagging, entity recognition and word-sense disambiguation. People usually solve such problems without difficulty or at least do a very good job. Linguistics may suggest labour-intensive ways of manually constructing rule-based systems. It is, however, the easy availability of large collections of texts that has made machine learning a method of choice for processing volumes of data well above the human capacity. One of the main purposes of text processing is all manner of information extraction and knowledge extraction from such large text. Machine learning methods discussed in this chapter have stimulated wide-ranging research in natural language processing and helped build applications with serious deployment potential.

NATURAL LANGUAGE PROCESSING AS A CHALLENGE FOR MACHINE LEARNING

The science and technology of the use and properties of human languages is generally known as Natural Language Processing (NLP). The name emphasizes, on the one hand, the natural origins of human spoken languages and, on the other, their role in computing. The discipline goes by several other mostly self-explanatory names, each bringing into focus a different key aspect: Computational Linguistics, Natural Language Understanding, Language Technology, Language Engineering, Text Processing, Text Analysis, etc. Disciplines that deal with speech, for example Speech Recognition and Speech Generation, may or may not be included in Natural Language Processing. Major classes of applications of theories, tools

DOI: 10.4018/978-1-60566-766-9.ch014

Table 1. Algorithms and corresponding learning problems discussed in some detail in this chapter.

Algorithm	Problem
Clustering by Committee	Learning concepts from text
Decision Trees	Classification of texts with features that are redundant (relevant but nor discriminative enough)
Deep Neural Networks	Combined semantic and probabilistic language modeling; multi-task feature learning
Finite-state automata	Stemming
Support Vector Machine	Classification of texts in high-dimensional feature spaces
Transformation-based learning	Part-of-speech tagging

and techniques developed in NLP include Machine Translation, Text Summarization, Text Mining and Information Extraction (IE). We refer the reader to (Mitkov, 2003) for an up-to-date description of NLP tasks and techniques. Feldman and Sanger (2007) discuss IE, visualization methods and probabilistic modelling, which often complement Machine Learning (ML) in language analysis. Sebastiani (2002) and Lopez (2008) present comprehensive overviews of ML applications to text categorization and translation, respectively.

Vocabulary, grammar, style – a few of many salient language dimensions – depend upon central dichotomies in natural language: spoken/written, edited/spontaneous, objective/subjective, formal/informal, and so on (Biber, 1988; Crystal, 2006). The nearly inexhaustible richness of language characteristics is "a challenge to those minds which seek ordered simplicity in the world, and at the same time a collectors' paradise" (Firth, 1936). The fast-increasing volume of readily available digital texts makes natural language not only a fertile ML research area (Daelemans, 2006), but also one of the most important data formats for ML applications (Shawe-Taylor and Christianini, 2004). The performance of ML algorithms is routinely compared on document topic classification, word-sense disambiguation and opinion analysis, to name just a few common NLP tasks.

This chapter presents ML applications to fundamental language-processing and linguistic problems: identify a word's part-of-speech or its meaning, find relations between words, and so on. Such problems, once solved, are a necessary component of many a higher-level language processing task, including text summarization, question answering and machine translation.

The next section of the chapter positions ML applications among technological tools for NLP. Next, we look at NLP from the ML perspective, focusing on such issues as the importance of annotated data, the characteristics of text data sets and the evaluation measures. We then discuss the nature of NLP problems solved by ML methods, which we propose to see as *nested learning* – based on the results of other learning. We show how Named-Entity Recognition, part-of-speech tagging and parsing contribute to learning in more advanced problems NLP problems. We look in some detail at ML applications to Word-Sense Disambiguation (WSD), one of the core NLP problems. Table 1 sums those algorithms and problems to which we have paid particular attention in this chapter. We discussed more algorithms and problems, but with fewer details.

Further in this handbook, the chapter "Machine Learning Applications in Mega-Text Processing" discusses in detail a few NLP tasks that emerged when text storage and exchange in electronic form became the norm.