Abstract

Memory-based Language Processing (MBLP) views language processing as being based on the direct reuse of previous experience rather than on the use of rules or other

1 Empirical Natural Language Processing

Although the origins of NLP are both logical and statistical, as in other disciplines of Artificial Intelligence, the knowledge-based approach has historically dominated this field. This has resulted in an emphasis on logical semantics for meaning representation, on the development of grammar formalisms (especially lexical unification grammars), and on the design of associated parsing methods and lexical representation and organization. Although these issues are central to NLP, they have also been introduced in the field, and have started to dominate it by the turn of the century.

From the early nineties onwards, empirical methods based on empirical algorithms have been introduced in the field. These methods are based on the direct reuse of previous experience rather than on the use of rules or other structures extracted from that experience. In such a framework, language acquisition is modeled as the storage of examples in memory, and language processing as analogical or similarity-based reasoning. For this reason, computational models of memory and similarity-based processes have been introduced in the field, and have started to dominate it by the turn of the century.

The emergence of empirical methods in NLP reflects an increased interest in the empirical methods that have been introduced in the field. Although these methods are based on empirical algorithms, they have also been introduced in the field, and have started to dominate it by the turn of the century.
This simple idea, and its many variants, has appeared regularly in work in artificial intelligence on the collection of experience, and the adaptation of the method that is described in the introduction of the next chapter. The reader may wonder why the process of learning is not adequately described by the experience of the system. The reason is that learning is the process of adapting the system to the environment, and it is this adaptation that is the subject of our study. In this chapter, we shall describe how the system adapts itself to the environment.

## Inspiration Sources

Memory-based learning and problem solving incorporate two principles: learning is the storage of experiences in memory, and solving a new problem is achieved by reusing solutions from similar previously solved problems. For an example in the language processing field: in the well-known prepositional phrase attachment problem, where it has to be decided by a language understander which verb or noun is modified by a particular prepositional phrase, traces of usage of earlier similar cases may help in the disambiguation. E.g., in a pizza with apple, the prepositional phrase with apple modifies pizza rather than eat because we have memory traces of similar expressions (e.g., pizza with anchovies, eat pizza with cheese, etc.). In a pizza with Eleni, other memory traces of similar sentence fragments such as crisps with Nicolas, and pizza with the boss would favor an adjective interpretation. For examples of the inductive process, see in the Maddow prepositional phrase database.

The simple idea, and its many variants, has appeared regularly in work in artificial intelligence on the collection of experience, and the adaptation of the method that is described in the introduction of the next chapter. The reader may wonder why the process of learning is not adequately described by the experience of the system. The reason is that learning is the process of adapting the system to the environment, and it is this adaptation that is the subject of our study. In this chapter, we shall describe how the system adapts itself to the environment.

## Inspiration Sources

Memory-based learning and problem solving incorporate two principles: learning is the storage of experiences in memory, and solving a new problem is achieved by reusing solutions from similar previously solved problems. For an example in the language processing field: in the well-known prepositional phrase attachment problem, where it has to be decided by a language understander which verb or noun is modified by a particular prepositional phrase, traces of usage of earlier similar cases may help in the disambiguation. E.g., in a pizza with apple, the prepositional phrase with apple modifies pizza rather than eat because we have memory traces of similar expressions (e.g., pizza with anchovies, eat pizza with cheese, etc.). In a pizza with Eleni, other memory traces of similar sentence fragments such as crisps with Nicolas, and pizza with the boss would favor an adjective interpretation. For examples of the inductive process, see in the Maddow prepositional phrase database.

The simple idea, and its many variants, has appeared regularly in work in artificial intelligence on the collection of experience, and the adaptation of the method that is described in the introduction of the next chapter. The reader may wonder why the process of learning is not adequately described by the experience of the system. The reason is that learning is the process of adapting the system to the environment, and it is this adaptation that is the subject of our study. In this chapter, we shall describe how the system adapts itself to the environment.

## Inspiration Sources

Memory-based learning and problem solving incorporate two principles: learning is the storage of experiences in memory, and solving a new problem is achieved by reusing solutions from similar previously solved problems. For an example in the language processing field: in the well-known prepositional phrase attachment problem, where it has to be decided by a language understander which verb or noun is modified by a particular prepositional phrase, traces of usage of earlier similar cases may help in the disambiguation. E.g., in a pizza with apple, the prepositional phrase with apple modifies pizza rather than eat because we have memory traces of similar expressions (e.g., pizza with anchovies, eat pizza with cheese, etc.). In a pizza with Eleni, other memory traces of similar sentence fragments such as crisps with Nicolas, and pizza with the boss would favor an adjective interpretation. For examples of the inductive process, see in the Maddow prepositional phrase database.
In intelligence, Psychology, Statistical Pattern Recognition, and Linguistics. This section describes the main inspiration sources for this type of algorithm. The next section situates the articles of this special issue within these traditions and in the context of related MBLP research.

Since Chomsky replaced the vague notions of analogy and induction existing in linguistics (in work of e.g. Saussure and Bloomfield) by the clearer and better operationalised notion of rule-based grammars, most mainstream linguistic theories, even the functionally and cognitively inspired ones, have assumed rules to be the only or main means to describe any aspect of language.

In contrast, the American linguist Royal Skousen (/1989, /1992/) argued for a specific operationalisation of the pre-Chomsky analogical approach to language and language learning (AML, Analogical Modeling of Language). He introduced a definition of analogy that is not based on rules and that treats all language data at the same level without making a distinction between regular instances (obeying the rules) and irregular instances (exceptions to the rules). To model language acquisition and processing, a database of examples of language use is searched looking for instances analogous to a new item, and extrapolating a decision for the new item from them. The linguistic motivation for this approach is (i) the fact that in actual language use there is not a clear-cut all-or-none distinction between regular and irregular cases, (ii) the simplicity of the analogical approach as opposed to rule discovery, and (iii) the adaptability of the approach as opposed to the static, rigid rule-based alternative. Remarkably, seen from the outside, such an analogical approach appears to be rule-governed, and therefore adequately explains linguistic intuitions as well.

The specific analogical algorithm employed by Skousen is available in a number of implementations. Current research attempts to solve the computational complexity problem (the algorithm is exponential in the number of attributes used to describe examples), and to apply the approach to a wide range of linguistic problems. The work has also been taken up as a psycholinguistically relevant explanation of human language acquisition and processing, especially as an alternative to dual route models of language processing (Eddington, /1998; Chandler, /1993; Derwing and Skousen, /1998). AML has also been used in computational linguistics. Jones (/1996) describes an application of AML in Machine Translation, and Daelemans, Gillis, and Durieux (/1997) compare AML to instance-based learning on a problem in computational phonology.

While AML is the most salient example of analogical approaches to languages in linguistics (and the most interesting from a computational linguistics point of view), other linguists outside the mainstream have proposed similar ideas. E.g. in the storage versus computation trade-off in models of linguistic processing, linguists like Bybee (/1988), and linguistic theories such as Cognitive Grammar (Langacker, /1991) claim an important role for examples (instances of language use); but they still presuppose rules to be essential for representing intuitions.

It is interesting to see that also in general psychology, in studies of human categorization, exemplar-based models often produce good fits of human behavior and errors (Smith and Medin, /1981; Nosofsky, /1986). These models assume that people represent a mental database of instances of categories in memory, and make categorization decisions based on the similarity of stimuli to these stored exemplars. They are contrasted with prototype, probabilistic or classical "rule-based" categorization models.
As far as the algorithms used in MBLP are concerned, nearest neighbor methods (k-nn), developed in statistical pattern recognition from the fifties onwards, have played an important inspirational role (e.g., Fix and Hodges, 1951; Cover and Hart, 1967). In these methods, examples labeled with their class are represented as points in an example space with as dimensions the numeric attributes used to describe the examples. A new example obtains its class by finding its position as a point in this space, and extrapolating its class from the k nearest points in its neighborhood. Nearness is defined as the reverse of Euclidean distance.

A very early citation nicely capturing the intuitive attraction of the k-nn approach is the following:

"This 'rule of nearest neighbor' has considerable elementary intuitive appeal and probably corresponds to practice in many situations. For example, it is possible that much medical diagnosis is influenced by the doctor's recollection of the subsequent history of an earlier patient whose symptoms resemble in some way those of the current patient."

(Fix and Hodges, 1952)

This literature has also generated many studies on methods for removing examples from memory either for efficiency (faster processing by removing unnecessary examples) or for accuracy (better predictions for unseen cases by removing badly predicting examples). See Dasarathy (1991) for a collection of fundamental papers on k-nn research.

However, while these methods may improve the performance of non-parametric statistical methods on the task at hand, they do not offer a solution to the broader problem of developing effective, generalizable, and robust non-parametric statistical methods for real-world applications. This has led to the development of k-nearest neighbor models as an alternative to conventional statistical methods for pattern recognition.

In some way, these models of the act of remembering, recall and recognition, correspond to practice in many situations. For example, in the case of near neighbor models, the memory of previous events is encoded in the relative position of an event in the sequence of events. However, while these methods may improve the performance of non-parametric statistical methods on the task at hand, they do not offer a solution to the broader problem of developing effective, generalizable, and robust non-parametric statistical methods for real-world applications.
of information, and the use of similarity-based reasoning as a smoothing method for low-frequency events. Especially the last property is an important theoretical issue. In language processing tasks, unlike other typical AI tasks, low-frequency events are pervasive. Due to borrowing, historical change, and the complexity of language, most data sets representing NLP tasks contain few regularities, and many subregularities and exceptions. It is impossible for inductive algorithms to reliably distinguish noise from exceptions, so non-abstracting lazy memory-based learning algorithms should be at an advantage compared to eager learning methods such as decision tree learning or rule induction: "forgetting exceptions is harmful." The usefulness of similarity for smoothing is discussed in Zavrel and Daelemans (1997) and Dagan, Lee and Pereira (1999). Daelemans, van den Bosch, and Zavrel (1999) provide empirical results and theoretical analysis supporting the "forgetting exceptions is harmful" hypothesis. Anbal van den Bosch (this volume) takes this analysis further by studying different methods of bottom-up abstraction from instances on a large range of NLP problems, and shows that it still holds in general. However, limited, careful abstraction using a notion of instance families (implemented in the FAMBL algorithm) can be used to prune memory without adverse effects on generalization accuracy on small data sets.

Memory-Based Computational Linguistics

Since the early nineties, we find several studies using nearest-neighbor techniques for solving NLP disambiguation problems, framed as classification problems. These tasks define modules that can play a role in different concrete applications. Each module is trained by collecting a set of examples with the required features for that module. E.g., morphosyntactic disambiguation (part-of-speech tagging) can be solved by representing information about the form of the word to be disambiguated and about the words in its immediate context as features, and the correct morphosyntactic category of the word in context as class. New words in context are assigned a class on the basis of the nearest neighbors. Such a tagger can then be used as a component of applications such as information extraction or translation, or several of such components can be combined to perform a more complex task, such as a parser.
3.2 Data-Oriented Parsing

Example-Based Machine Translation

This approach involves translating a sentence from one language to another by finding similar sentences in a corpus of pre-translated examples and using those translations as a guide. It is particularly useful for dealing with languages with flexible word order and a large number of possible sentence structures. The approach is based on the idea that similar sentences should have similar translations, which can be used to predict the translation of new sentences. This approach has been successfully applied to a wide range of languages, including those with complex grammatical structures.

References:
- Nagao (1984) proposed an approach to machine translation, which is
- Because of the huge space of possible sentences to be translated, and the cost of collecting and searching large amounts of pre-translated EBMT examples, EBMT systems are mostly hybrid, and contain rule-based, statistical, and case-based components. Andy Way (1999) proposes an approach to machine translation, which is

Other approaches to machine translation include statistical models, neural networks, and hybrid systems. Each approach has its own advantages and disadvantages, and the choice of approach depends on the specific application and the available resources.

3.3 Other Approaches

This section covers a range of other approaches to machine translation, including rule-based systems, statistical models, and neural networks. Each approach has its own strengths and weaknesses, and the choice of approach depends on the specific application and the available resources.

Rule-Based Systems

Rule-based systems use a set of explicit rules to determine the correct translation of a sentence. These rules are typically based on syntactic and semantic considerations, and they are often manually designed by linguists. Rule-based systems are good at handling complex sentences with irregular word order, but they can be difficult to develop and maintain.

Statistical Models

Statistical models use statistical algorithms to learn the best translation of a sentence. These models are trained on a large corpus of pre-translated examples, and they learn to predict the translation of new sentences based on the statistical properties of the training data. Statistical models are good at handling large amounts of data, but they can be less accurate than rule-based systems for complex sentences.

Neural Networks

Neural networks use a set of interconnected nodes (neurons) to learn the best translation of a sentence. These networks are trained using a large corpus of pre-translated examples, and they learn to predict the translation of new sentences based on the connections between the neurons. Neural networks are good at handling large amounts of data, and they can be more accurate than rule-based systems for complex sentences.
The special issue has been made possible by grants from FWO (Belgium) and NWO and the authors are grateful for their cooperation and their critical comments. They work on this collection. Many thanks also to the editors of the special issue for their help and support and encouragement of the workshop for their support and encouragement.

Translation memories, arguably the most successful commercial approach to machine-aided translation to date, are based on the memory-based framework: large amounts of documents aligned with their translations are stored, and a possible translation for a new sentence is searched using a fuzzy string matching technique.

Analogy and Similarity

The concept of analogy implicit in memory-based classification approaches, based on the nearest-neighbor metaphor, is simple and computationally efficient for many information-processing tasks. In this special issue, we have collected memory-based analogy approaches characterized by the following three features:

1. They have a fixed number of features.
2. They have a fixed number of classes.
3. They have a fixed number of examples.

The concept of analogy implicit in memory-based classification approaches based on a fixed number of features, classes, and examples is simple and computationally efficient for many information-processing tasks.
Table 1: Themes and tasks addressed by the articles in this special issue.


Gazdar/, G./, and Mellish/, C./, 1989, Natural Language Processing in LISP (Reading/, MA/: Addison-Wesley/).


Jones/, D./, 1996, Analogical Natural Language Processing (London/: UCL Press/).


Koleda/, J./, 1993, Case-Based Reasoning (London/: PTR Prentice Hall/).

Kolodner/, J./, 1993, Case-Based Reasoning (San Mateo/: Morgan Kaufmann/).
