TALN 2003
Using decision trees to learn lexical information in a linguistics-based NLP system

Marisa Jiménez and Martine Pettenaro

Microsoft Research
One Microsoft Way
Redmond, WA 98052, USA
marialj@microsoft.com
martinep@microsoft.com

Résumé – Abstract

Nous décrivons dans cet article l'utilisation d'arbres décisionnels pour l'acquisition d'informations lexicales et l'enrichissement de notre système de traitement automatique des langues naturelles (NLP). Notre approche diffère d'autres projets d'apprentissage automatique en ce qu'elle repose sur l'exploitation d'un système d'analyse linguistique profonde. Après l'introduction de notre sujet nous présentons l'architecture de notre module d'apprentissage lexical. Nous présentons ensuite une situation d'apprentissage lexical effectué en utilisant des arbres décisionnels; nous apprenons quels verbes prennent un sujet humain en espagnol et en français.

This paper describes the use of decision trees to learn lexical information for the enrichment of our natural language processing (NLP) system. Our approach to lexical learning differs from other approaches in the field in that our machine learning techniques exploit a deep knowledge understanding system. After the introduction we present the overall architecture of our lexical learning module. In the following sections we present a showcase of lexical learning using decision trees: we learn verbs that take a human subject in Spanish and French.

Keywords – Mots Clés

Apprentissage lexical, apprentissage automatique, arbres décisionnels, dictionnaires automatiquement appris.
Lexical learning, machine learning, decision trees, learned dictionaries.

1 Introduction

In this paper we describe the use of a particular machine learning technique, decision trees (DTs), to acquire lexical information in a broad-coverage linguistics-based natural language processing (NLP) system.

The manual encoding of lexical information into an online dictionary is costly and time-consuming. Therefore, the NLP community has used various machine learning (ML) techniques to automatically learn lexical information (Atwell et al. (1994), Teufel (1995), Stevenson and Merlo (1997), Van Halteren et al. (1998), Brill and Wu (1998), Schulte im Walde (1998), Stevenson et al. (1999), Zavrel and Daelemans (2000), Van Halteren et al. (2001), among others). Most approaches consist of ML techniques that make use of linear

context (n-grams extracted from text) and that are not integrated in a rich NLP system. By contrast we use a particular ML technique, DTs, that exploits a deep knowledge understanding system to learn lexical information, which in turn immediately benefits our system. Our goal was to prove that DTs offer many advantages as a lexical information acquisition technique:

- They are much faster than manual encoding.
- They are more flexible than purely linguistic heuristics. For example, they allow us to test many linguistic features at once. Also, they can be called at run-time from any module of our system to make lexical predictions.
- DT tools suit our needs better than support vector machines (SVMs): they permit thorough error analysis and allow us to use probability distributions.
- They are a useful classification tool for computational lexicographers; they unveil useful linguistic patterns that may not be discerned by the native speaker/linguist.

In section 2 we present an overview of the architecture of the lexical learning module which allows us to exploit DTs. In the remainder of the paper we present a showcase of lexical learning using DTs; we learn verbs that take a human subject in Spanish and French.

2 Lexical learning in our system

Our NLP system uses various learning techniques to acquire lexical information. This new lexical information supplements the knowledge already contained in our monolingual dictionaries. For example we can learn unknown words encountered in text, corpus-specific words or collocations and various features pertaining to them, and part-of-speech probabilities.

Our system has two main techniques in place to perform lexical learning. The first technique consists of linguistic rules that allow us to acquire lexical information at any stage of sentence processing: either morphological, named entity, syntactic, or logical-form processing may be adequate depending on the type of information we are interested in; for example, we might want to learn the feature "count noun" through plural nouns found in the data. This information is stored in learned dictionaries that supplement the general dictionary (see Pentheroudakis (2001), and Wu et al. (2002)). Any subset of learned dictionaries can be used depending on corpus-specific needs, and their entries can also be permanently merged with the main dictionary if desired.

The second technique consists of lexical learning through the use of DTs. The tools that we use to build our DTs are borrowed from the publicly available WinMine toolkit (Chickering, n.d.), developed at Microsoft Research. The WinMine tools automatically split the data into training and testing (70/30), and produce several DT models at different levels of granularity.

DTs provide a classification of selected features and rank their relative importance in predicting the target feature we are trying to learn (for example, "takes human subject"). They also provide a probability distribution over all possible target value paths, i.e., over all possible combinations found in the data of the various features with a value (true or false) of the target feature.

Lexical learning via DTs proceeds in three stages:

- Unsupervised annotation of the target feature for relevant data points in a selected corpus, and linguistic feature extraction from the parsed sentences in which each data point occurs.
- Build DT models using the results from the previous step. The task is to classify and assign probabilities to the contexts in which each data point occurs.
- Dynamically add new linguistic information to a learned dictionary, based on the

predictions made by the best model. Select a probability distribution threshold.

3 Showcase: learning verbs that take a human subject in Spanish and French

In this showcase, our goal was to learn which verbs take a human subject in French and Spanish. The motivation for this project was to improve the translation of personal pronouns in our French->English and Spanish->English machine translation (MT) systems. We were focusing on improvements in the translation of: 1) the French Canadian Hansard, 2) Spanish technical manuals.

English generation needs more information than just the French and Spanish pronoun forms, if available, to generate subject pronouns correctly. Unlike English, Spanish can omit subject pronouns. In our English technical corpora, we allow two possible translations for the subject of the Spanish verb *escribe* in (1a): either *it* or *you*, which we illustrate in (1b) and (1c), respectively:

(1a) *Por ejemplo, si* ***escribe*** *la expresión París, el programa la mostrará como sigue.*
(1b) *For example, if* ***it enters*** *the expression Paris, the program will display it as follows.*
(1c) *For example, if* ***you enter*** *the expression Paris, the program will display it as follows.*

Our English generation component has no way of preferring one translation over another for the non-overt subject, unless our system is provided with the information that the Spanish verb typically subcategorizes for a human subject. This information would allow English to generate *you enter* as the translation of *escribe* rather than the default *it enters*.

In French the issue centers around the pronoun *il*,[1] which can be human or non-human. If it is human *il* is translated in English as *he*, while if it is non-human, it should be translated as *it*. In (2a) we provide a sample French sentence with two instances of *il*; in (2b) and (2c) we provide our original default translation and the improved translation, respectively:

(2a) ***Il a dit*** *que, lorsque les sénateurs seront élus,* ***il sera*** *presque impossible de renégocier la répartition des sièges.*
(2b) ***It has said*** *that when the senators will be elected,* ***it will be*** *almost impossible to renegotiate the distribution of the seats.*
(2c) ***He has said*** *that when the senators will be elected,* ***it will be*** *almost impossible to renegotiate the distribution of the seats.*

These facts led us to run our cross-linguistic experiment. In order to learn which verbs take a human subject in French or Spanish, we wanted to use DTs since they are well known as a classification tool. Our goal was to classify the contexts in which verbs taking a human subject appear, and to use the prediction of our DT models to create dictionary records containing these same verbs with the human subject feature.

3.1 Data annotation and feature extraction

Our Spanish and French systems contained little or no information about whether verbs subcategorize for a human subject. The French monolingual dictionary did not contain any verbs bearing the human subject (*HSubj*) tag. Spanish had a handful.

Given this lack of information, we decided to automatically tag the verbs in our data with the

[1] We ignored the pronoun *elle* for this experiment, as it is much rarer in our corpus.

target feature (*HSubj*). The main advantage of automatically tagging corpora is that one can use much more training data. We gathered 405,810 sentences from the French Canadian Hansard (record of parliamentary debates) and 350,516 sentences from the Spanish *Encarta* encyclopedia.[2] In these sentences we assigned the *HSubj* tag to verbs whose subject was identified as human by our system; this operation relied heavily on the human feature which is assigned to relevant nouns in our dictionaries. We excluded sentences containing verbs with subjects whose head was ambiguous regarding the human feature, or was not found in our dictionary.

Automatic tagging is obviously more susceptible to errors than manual tagging. We relied on our broad-coverage parsers of Spanish and French for the correct identification of subject noun phrases (NPs); parse accuracy is not always perfect, but these parsers have been put to the test in the past. We also trusted our monolingual French and Spanish dictionaries with information about human nouns. Furthermore, we reviewed a random sample of the tags to make sure that our automatic tagging scheme was robust.

During the data annotation procedure we also automatically extracted 166 features for Spanish and 176 features for French from each of the sentences in the data sets; in each case a particular feature was assigned a 1 if it was found to be true or a 0 if it was false. We extracted the following groups of features:

- Morphological, syntactic and semantic features of the main verb.
- Morphological, syntactic and semantic features of the subject of the verb.
- Syntactic features of the clause.

3.2 Decision tree results

We built DT models using the WinMine toolkit on the annotated data described above.

In Table 1 we provide the results for the best DT model for French, and in Table 2 the results for Spanish.

Precision represents the number of correct predictions for the positive or negative values of the target feature (*Hsubj* or *NoHSubj*), divided by all predictions for that value. *Recall* calculates the number of correct predictions for one value of the target feature, divided by the number of verbs bearing that value in the actual data. The *F-measure* is a combination of precision and recall. The *baseline* represents the percentage of times the correct prediction for the most common value of the target feature (*NoHSubj*) would occur if we did not run the DT. *Accuracy* represents our overall correct predictions, whether *Hsubj* or *NoHSubj*, divided by all values of the target feature in the data.

Out of the 176 and 166 features extracted for French and Spanish respectively, 87 features were determined by WinMine to be relevant for French, while 98 were found to be relevant for Spanish. Examples of the most relevant features across languages follow:

- Features of the subject: definiteness, gender, number, upper case, proper name, profession, title, possessor, presence of a complementizer or apposition, location...
- Features of the verb: polite form, speech act, 1st and 2nd person, perfect tense, governing preposition, reflexivity, transitivity, object control...
- Features of the clause: infinitive clause, main clause, main clause with dependent clauses, dependent clause, topic…

Manual inspection of the DTs confirmed the quality of the DT results. We partly attribute the difference in the results between French and Spanish to the peculiarities of the data: we also experimented with French using the French *Encarta* encyclopedia, and did not obtain as good

[2] We did not use technical manuals during the learning phase because they did not contain enough overt information about human subjects.

results as with Hansard. Furthermore, Spanish had less training data because of its covert subjects.[3]

	HSubj	NoHSubj	Overall
Precision	95.41%	92.49%	
Recall	91.25%	96.08%	
F-measure	93.28%	94.25%	
Baseline			52.86%
Accuracy			93.80%

Table 1: French results

	HSubj	NoHSubj	Overall
Precision	86.16%	87.52%	
Recall	73.00%	94.17%	
F-measure	79.04%	90.72%	
Baseline			66.79%
Overall Accuracy			87.14%

Table 2: Spanish results

3.3 Building learned dictionaries with human subject tags on verbs

The next step in the experiment consisted in building learned dictionaries for French and Spanish which would contain verb records with the *HSubj* tag.

We invoked the best DT models for Spanish and French while parsing their respective corpora. We used a lexicalization function to add a new verbal record with the *HSubj* tag to the learned dictionary only if the predictions of the DT model were above a .75 probability threshold. We set a high threshold to minimize errors.

This process resulted in a French learned dictionary containing 2,415 *HSubj* verbs, and a Spanish learned dictionary with 1,975 *HSubj* verbs.

3.4 Evaluation of the learned dictionaries

In the last step of the project we evaluated the impact of the learned dictionaries in our French ->English and Spanish->English MT systems. Again, our goal was to test whether the translations of covert subjects in Spanish and of the French pronoun *il* could be improved by using the *HSubj* learned dictionaries.

French ->English

We took a random selection of 500 French sentences containing the subject pronoun *il* from the Hansard data. To create a baseline we translated these sentences without using the *HSubj* learned dictionary. We counted the number of times that *il* was correctly translated in English. The baseline accuracy was 53.15%, i.e., the English generation's default translation *it* was correct 53.15% of the time. Then we ran the 500 sentences again while using the learned dictionary, with generation taking the *HSubj* into account; we reached an accuracy of 84.84%.

Spanish->English

We followed the same strategy as for French-> English. We gathered a random sample of 500 sentences from technical manuals containing null subjects. To create a baseline we counted the number of times the null subject was translated correctly into English without the help of the learned dictionary. In order to perform a fair evaluation, we commented out English

[3] We obtained 318,675 *HSubj* data points for French, and only 116,595 for Spanish (although French only had roughly 50,000 sentences more than Spanish).

generation code that was already in place to guess the correct generation of some covert subject pronouns. The baseline was 31.19%, i.e., the default translation *it* was correct 31.19% of the time. The rest of the time, for those technical manuals, *you* would have been a more appropriate translation (in non-imperative instructions given to the reader). Then we processed the 500 sentences again, this time using the learned dictionary, and obtained an accuracy of 83.37%.

4 Conclusion

In this paper we described our use of DTs to acquire lexical information. We proved that they constitute a useful tool for lexical learning when used within the framework of a deep knowledge understanding system. We showed that by creating learned dictionaries using DTs we benefited our NLP system; for example we were able to improve the translation of personal pronouns in our French->English and Spanish-> English MT systems. In addition, not only can we create references like dictionaries through DTs, but we can also call DTs from anywhere in our system to make predictions about new lexical items at run-time.

References

Atwell, E., J. Hughes, and C. Souter (1994). *Amalgam: Automatic Mapping among Lexico-grammatical Annotation Models*. Technical report, Internal Paper, CCALAS, Leeds University.

Brill, E. and J. Wu (1998). *Classifier Combination for Improved Lexical Disambiguation*. COLING-ACL'98. Montreal, Canada.

Chikering, D. Max (n.d.) *WinMine Toolkit Home Page*:

http://www.research.microsoft.com/~dmax/WinMine/Tooldoc.htm.

Pentheroudakis, J. (2001). *Lex Rules!*. Technical report, Microsoft Research.

Schulte im Walde, S., 1998. *Automatic Semantic Classification of Verbs according to their Alternation Behaviour*. AIMS Report 4(3), IMS, Universität Stuttgart.

Stevenson, S. and P. Merlo (1997). *Lexical Structure and Processing Complexity*. Language and Cognitive Processes, 12(1-2):349-399.

Stevenson, S., P. Merlo, N. Karaeva, and K. Whitehouse (1999). *Supervised Learning of Lexical Semantic Verb Classes using Frequency Distributions*. Procs of SigLex '99, College Park, Maryland.

Teufel, S. (1995). *A Support Tool for Tagset Mapping*. Procedings of the Workshop SIGDAT (EACL95).

Van Halteren, H., J. Zavrel, and W. Daelemans (1998). *Improving Data Driven Wordclass Tagging by System Combination*. Proceedings of ACL-COLING'98, Montreal, Canada.

Van Halteren, H., J. Zavrel, and W. Daelemans (2001). *Improving Accuracy in NLP through Combination of Machine Learning Systems*. Computational Linguistics 27 (2), pp. 199-230.

Wu, A., J. Pentheroudakis, and Z. Jiang (2002). *Dynamic Lexical Acquisition in Chinese Sentence Analysis*. Proceedings of COLING 2002, Taipei, Taiwan.

Zavrel, H.J. and W. Daelemans (2000). *Bootstrapping a Tagged Corpus through Combination of Existing Heterogeneous Taggers*. International Conference on Language Resources and Evaluation, Athens, Greece.