Entropy Guided Transformation Learning

Cícero Nogueira dos Santos[1,2] and Ruy L. Milidiú[1]

[1] Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro
Rio de Janeiro, Brazil

[2] Mestrado em Informática Aplicada, Universidade de Fortaleza – UNIFOR
Fortaleza, Brazil

cnogueira@unifor.br,milidiu@inf.puc-rio.br

Abstract. This work presents Entropy Guided Transformation Learning (ETL), a new machine learning algorithm for classification tasks. It generalizes Transformation Based Learning (TBL) by automatically solving the TBL bottleneck: the construction of good template sets. We also present ETL Committee, an ensemble method that uses ETL as the base learner. The main advantage of ETL is its easy applicability to Natural Language Processing (NLP) tasks. Its modeling phase is quick and simple. It only requires a training set and a naive initial classifier. Moreover, ETL inherits the TBL flexibility to work with diverse feature types. We describe the application of ETL to four language independent NLP tasks: part-of-speech tagging, phrase chunking, named entity recognition and semantic role labeling. Overall, we apply it to thirteen different corpora in six different languages: Dutch, English, German, Hindi, Portuguese and Spanish. Our extensive experimental results demonstrate that ETL is an effective way to learn accurate transformation rules. Using a common parameter setting, ETL shows better results than TBL with handcrafted templates for the four tasks. For the Portuguese language, ETL obtains state-of-the-art results for all tested corpora. Our experimental results also show that ETL Committee improves the effectiveness of ETL classifiers. Using the ETL Committee approach, we obtain state-of-the-art competitive performance results in the thirteen corpus-driven tasks. We believe that by avoiding the use of handcrafted templates, ETL enables the use of transformation rules to a greater range of NLP tasks.

1 Introduction

Since the last decade, Machine Learning (ML) has proven to be a very powerful tool to help in the construction of Natural Language Processing (NLP) systems, which would otherwise require an unfeasible amount of time and human resources. Transformation Based Learning (TBL) is a ML algorithm introduced by Eric Brill [1] to solve NLP tasks. TBL is a corpus-based, error-driven approach that learns a set of ordered transformation rules which correct mistakes of a baseline classifier. It has been used for several important NLP tasks, such as part-of-speech tagging, phrase chunking, parsing, named entity recognition and semantic role labeling.

TBL rules must follow patterns, called templates, that are meant to capture the relevant feature combinations. TBL templates are handcrafted by problem experts. Its quality strongly depends on the problem expert skills to build them. Even when a template set is available for a given task, it may not be effective when we change from a language to another. When the number of features to be considered is large, the effort to manually create templates is extremely increased, becoming sometimes infeasible. Hence, the human driven construction of good template sets is a bottleneck on the effective use of the TBL approach.

In this work, we present Entropy Guided Transformation Learning (ETL), a new machine learning algorithm for classification tasks. ETL generalizes TBL by automatically solving the fourteen years old TBL bottleneck: the construction of good template sets. ETL uses the information gain in order to select the feature combinations that provide good template sets. ETL provides an effective way to handle high dimensional features. It also enables the inclusion of the *current classification* feature in the generated templates. In this work, we also present ETL Committee, an ensemble method that uses ETL as the base learner.

ETL can be easily applied to NLP tasks. Its modeling phase is quick and simple. It only requires a training set and a naive initial classifier. Moreover, ETL inherits the TBL flexibility to work with diverse feature types. In order to assess the robustness and predictive power of the ETL strategy, we apply it to four language independent NLP tasks: Part-of-Speech (POS) Tagging, Phrase Chunking (PCK), Named Entity Recognition (NER) and Semantic Role Labeling (SRL). These tasks have been considered fundamental for more advanced NLP applications. Overall, we apply ETL to thirteen different corpora in six different languages: Dutch, English, German, Hindi, Portuguese and Spanish.

In Table 1, we summarize, for each corpus, the performance of both ETL and the state-of-the-art system[3]. The performance measure for the POS tagging task is accuracy. For the other three tasks, the performance measure is the $F_{\beta=1}$. In Table 1, the best observed results are in bold. Using the ETL approach, we obtain competitive performance results for the thirteen corpora. For each one of the tasks, ETL shows better results than TBL with handcrafted templates. For the Portuguese language, ETL obtains state-of-the-art results for the four tested corpora. ETL Committee improves the effectiveness of ETL classifiers in all cases and achieves state-of-the-art results for six corpora.

ETL and ETL Committee are introduced in two papers presented by the authors in two main international conferences on NLP. In Milidiú et al. [3], a paper presented in the *46th Annual Meeting of the Association for Computational Linguistics - ACL 2008*, we introduce ETL and show its application to language independent PCK. In dos Santos et al. [4], a paper presented in the *11th International Conference on Intelligent Text Processing and Computational Linguistics – CICLing 2010*, we present a detailed description of ETL Committee and apply it to Text Chunking, NER and SRL.

The ETL effectiveness for NLP tasks can be also verified in a series of papers published in journals and in the proceedings of important international

[3] References to the state-of-the-art systems can be found in dos Santos [2].

Table 1. System performances.

Task	Corpus	Lang.	State-of-the-art		ETL	
			System	Performance	Single	Committee
POS	Mac-Morpho	POR	TBL	96.60	96.75	**96.94**
	Tycho Brahe	POR	TBL	96.63	96.64	**96.72**
	Brown	ENG	TBL	96.67	96.69	**96.83**
	TIGER	GER	TBL	96.53	96.57	**96.68**
PCK	SNR-CLIC	POR	TBL	87.71	88.85	**89.58**
	Ramshaw & Marcus	ENG	SVM	**94.22**	92.80	93.29
	CoNLL-2000	ENG	SVM	**94.12**	92.28	93.27
	SPSAL-2007	HIN	HMM + CRF	**80.97**	78.53	80.44
NER	HAREM	POR	CORTEX	61.57	61.32	**63.56**
	SPA CoNLL-2002	SPA	AdaBoost	**79.29**	76.28	77.46
	DUT CoNLL-2002	DUT	AdaBoost	**77.05**	74.18	75.44
SRL	CoNLL-2004	ENG	SVM	**69.49**	63.37	67.39
	CoNLL-2005	ENG	AdaBoost	**75.47**	70.08	72.23

conferences. In dos Santos & Milidiú [5], we present through a book chapter a detailed description of the ETL algorithm and apply it to Text Chunking, NER and SRL. In Milidiú et al. [6], a paper published in the *Journal of the Brazilian Computer Society*, we present ETL models that obtain state-of-the-art results for three Portuguese language processing tasks: POS tagging, PCK and NER. In dos Santos et al. [7], a paper presented in the *PROPOR'2008* conference, we show a state-of-the-art ETL-based Portuguese POS tagger. Furthermore, ETL allowed the use of transformation rules for tasks where TBL has not been applied before. In the *PROPOR'2010* conference, we present an application of ETL to Portuguese clause identification [10]. In the *STIL'2009* event, we published two papers on applications of ETL to English clause identification [8] and Portuguese dependency parsing [9]. Additionally, in a very short period of time, ETL allowed the creation of F-EXT-WS [11], a web based Natural Language Processor. For the Portuguese language, this web service provides the processing of various tasks such as POS tagging, NP chunking and NER. Finally, due to its potential to allow the quick development of information extractors, ETL has won the *Prêmio Solução Rio Info 2009*, a Brazilian contest that annually awards the best PhD dissertation on Information Technology.

The remainder of the paper is organized as follows. In Sections 2 and 3, we briefly describe ETL and ETL Committee strategies.In Section 4, we summarize our experiments. Finally, in Section 5, we present some concluding remarks.

2 Entropy Guided Transformation Learning

Entropy Guided Transformation Learning employs an *entropy guided template generation* approach, which uses Information Gain (IG) in order to select the feature combinations that provide good template sets [5]. The learning strategy

requires two inputs: a training corpus and a baseline classifier. The template generation step is performed using Decision Tree (DT) induction, which employs IG for feature selection. The ETL algorithm is illustrated in the Figure 1. It can also be formulated as follows:

1. A set of rule templates is generated by decomposing a DT. The DT is induced from the training corpus. For each DT node u, a template is created by combining the features in the path from the root node to u.
2. The baseline classifier is applied to the training set.
3. *Repeat*:
   (a) The current classification is compared with the correct one and, whenever a classification error is found, all the rules that can correct it are generated by instantiating the templates. Usually, a new rule corrects some errors, but also generates some other errors by changing correctly classified samples.
   (b) The rules' scores (repaired errors - created errors) are computed.
   (c) *Stop*, if there is not a rule whose score is above a given threshold.
   (d) The best scoring rule is selected, stored in the list of learned rules and applied to the training set.

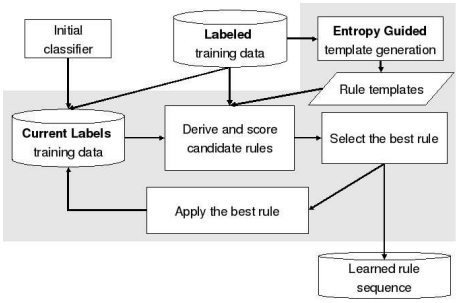

Fig. 1. Entropy Guided Transformation Learning.

Observe that steps 2 and 3 correspond to the TBL algorithm. ETL enables the use of high dimensional features by appling a simple preprocessing phase at step 1. Using the true class information in the DT trainig, ETL also enables the inclusion of the *current classification* feature in the generated templates. A detailed description of ETL can be found in dos Santos [2].

3 ETL Committee

ETL Committee is an ensemble method that uses ETL as the base learner. It combines the main ideas of Bagging [12] and Random Subspacesv [13]. From Bagging, it borrows the bootstrap sampling method. From Random Subspaces, it uses the feature sampling idea. In the ETL Committee training, template sampling is used to speed up the training phase.

3.1 ETL Committee Training phase

Given a labeled training set $\mathcal{T}$, the ETL Committee algorithm generates L ETL classifiers using different versions of $\mathcal{T}$. In Figure 2, we detail the ETL Committee training phase. The creation of each classifier is independent from the others. Therefore, the committee training process can be easily parallelized. In the creation of a classifier c, the first step consists in using *bootstrap sampling* to produce a bootstrap replicate $\mathcal{T}^{'}$ of the training set $\mathcal{T}$. Next, *feature sampling* is applied to $\mathcal{T}^{'}$, generating the training set $\mathcal{T}^{''}$. Finally, in the *ETL training* step, a rule set is learned using $\mathcal{T}^{''}$ as a training set. These steps are detailed in [2].

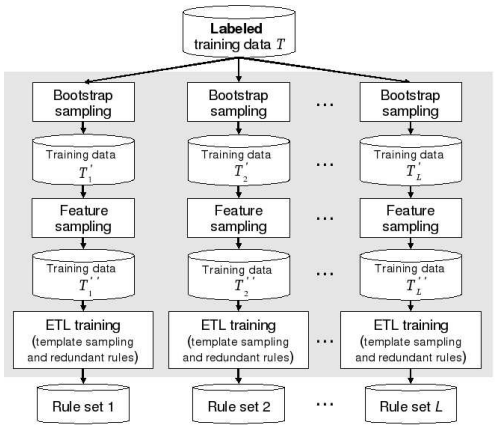

Fig. 2. ETL Committee training phase.

3.2 ETL Committee Classification phase

When classifying new data, each transformation rule set is independently applied to the input data. For each data point, each ETL model gives a classification, and we say the model "votes" for that class. The final data point classification is computed by majority voting.

A drawback of ETL Committee, as well as the other ensemble methods, is that it increases the classification time. However, this process can be easily parallelized, since the application of each rule set is independent from the others.

4 Experiments

This section presents the experimental setup and results of the application of ETL and ETL Committee to four NLP tasks: Part-of-Speech (POS) Tagging, Phrase Chunking (PCK), Named Entity Recognition (NER) and Semantic Role Labeling (SRL). ETL and ETL Committee results are compared with the state-of-the-art system results for each corpus.

4.1 Machine Learning Modeling

The four tasks are modeled as token classification problems. Which means that, given a text, the learned system must predict a class label for each token.

We use the following ETL and ETL Committee common parameter setting in our experiments with the four tasks. The parameters are empirically tuned using the training and development sets available for the NER and SRL tasks.

ETL Single: we use a context window of size seven. We use templates which combine at most six features. Therefore, when extracting templates from DTs, the extraction process examines only the six first DT levels. We let the ETL algorithm learn rules whose score is at least two.

ETL Committee: for the ETL Committee, in the *bootstrap sampling* step, we use sentences as sampling units for bootstrapping. We set the ensemble size to 100. In the *feature sampling* step, we randomly sample 90% of the features for each classifier. In the *ETL training* step, we let the ETL algorithm to learn the largest rule set possible. We use 50 as the default number of templates to be sampled in the creation of each classifier.

BLS: the initial classifiers, or baseline systems (BLS), use simple statistics of the training data. The details are available in dos Santos [2].

4.2 Part-of-Speech Tagging

Part-of-Speech (POS) tagging is the process of assigning a POS or another lexical class marker to each word in a text. POS tags classify words into categories, based on the role they play in the context in which they appear. The POS tag is a key input feature for NLP tasks like PCK and NER.

We apply ETL to four different POS tagged corpora in three different languages. The selected corpora are: Mac-Morpho, a Portuguese language corpus; Tycho Brahe, a Portuguese language corpus; TIGER, a German language corpus; and Brown, an English language corpus.

Our POS tagging modeling approach follows the two stages strategy proposed by Brill [1]. The first stage, the *morphological*, classifies the unknown words using morphological information. The second stage, the *contextual*, classifies the known and unknown words using contextual information. We use ETL and ETL Committee for the contextual stage only, since the morphological stage uses trivial templates.

In Table 1, we summarize the results for the POS tagging corpora. As we can see, the ETL system achieves state-of-the-art performance for the four corpora. ETL Committee slightly improves the ETL accuracy.

4.3 Phrase Chunking

Phrase Chunking (PCK) consists in dividing a text into non-overlapping phrases. It provides a key feature that helps on more elaborated NLP tasks such as parsing and SRL. We apply ETL to four different PCK corpora in three different

languages. The selected corpora are: SNR-CLIC, a Portuguese NP chunking corpus; Ramshaw & Marcus (R&M), an English base NP chunking corpus; CoNLL-2000, an English PCK corpus; and SPSAL-2007, a Hindi PCK corpus.

The experimental results for the PCK task are summarized in Table 1. ETL achieves state-of-the-art performance for the Portuguese and the Hindi corpora. For the other two, R&M and CoNLL-2000, the ETL system achieves state-of-the-art competitive results. ETL Committee significantly improves the ETL results for the four corpora. For the CoNLL-2000 Corpus, ETL Committee reduces the $F_{\beta=1}$ error by 13% when compared to a single ETL model.

4.4 Named Entity Recognition

Named Entity Recognition (NER) is the problem of finding all proper nouns in a text and to classify them among several given categories of interest. Usually, there are three given categories: Person, Organization and Location. Our NER approach follows the two stages strategy proposed in [1] for POS tagging.

We apply ETL to three different NER corpora in three different languages. The selected corpora are: HAREM, a Portuguese NER corpus; SPA CoNLL-2002, a Spanish NER corpus; and DUT CoNLL-2002, a Dutch NER corpus.

In Table 1, we summarize the results for the NER task. For the three corpora, the ETL system achieves state-of-the-art competitive results. Similarly to the PCK task, ETL Committee significantly improves the ETL results.

4.5 Semantic Role Labeling

Semantic Role Labeling (SRL) is the process of detecting basic event structures such as *who* did *what* to *whom*, *when* and *where*. More specifically, for each predicate of a clause, whose head is typically a verb, all the constituents in the sentence which fill a semantic role of the verb have to be recognized.

We evaluate the performance of ETL over two English language corpora: CoNLL-2004 and CoNLL-2005. These two corpora were used in the CoNLL shared task of the years 2004 and 2005, respectively. Since our purpose is to examine ETL and ETL Committee performance for a complex task, we do not use the full parsing information in our SRL experiments.

The experimental results for the SRL task are summarized in Table 1. The ETL system achieves regular results for the two corpora. However, ETL Committee improves the ETL results for the two corpora. When compared to the single ETL systems, ETL Committee reduces the $F_{\beta=1}$ error by 11% and 7% for the CoNLL-2004 and CoNLL-2005 Corpus, respectively. We believe that the error reduction is larger for the CoNLL-2004 Corpus because this corpus is smaller.

5 Conclusions

Entropy Guided Transformation Learning is a machine learning algorithm that generalizes TBL. ETL solves the fourteen years old TBL bottleneck: the construction of good template sets. In this work, we also present ETL Committee,

a new ensemble method that uses ETL as the base learner. We summarize the experimental design and results of the application of ETL to four NLP tasks: POS Tagging, PCK, NER and SRL. Overall, we apply it to thirteen different corpora in six different languages: Dutch, English, German, Hindi, Portuguese and Spanish. Our extensive experimental results support the hypotheses that ETL is an effective way to learn accurate transformation rules for NLP tasks. For each one of the tasks, ETL shows better results than TBL with handcrafted templates.

Our experimental results indicate that ETL Committee significantly outperforms single ETL models. It is worth to mention that the proposed ensemble method improves the ETL effectiveness without any additional human effort. Moreover, it is particularly useful when dealing with large feature sets.

For the Portuguese language, ETL achieves the best reported results in all tested corpora. Moreover, in a very short period of time, it allowed the creation of F-EXT-WS [11], a web based Natural Language Processor. This web service provides the processing of various tasks for the Portuguese language, such as: POS tagging, noun phrase chunking and NER.

References

1. Brill, E.: Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. Comp. Linguistics **21**(4) (1995) 543–565
2. dos Santos, C.N.: Entropy Guided Transformation Learning. PhD thesis, Pontifícia Universidade Católica do Rio de Janeiro (2009)
3. Milidiú, R.L., dos Santos, C.N., Duarte, J.C.: Phrase chunking using entropy guided transformation learning. In: Proceedings of ACL2008, Columbus, Ohio (2008)
4. dos Santos, C.N., Milidiú, R.L., Crestana, C.E.M., Fernandes, E.R.: Etl ensembles for chunking, ner and srl. In: Proceedings of the 11th CICLing. (March 2010) 100–112
5. dos Santos, C.N., Milidiú, R.L.: Entropy Guided Transformation Learning. In: Foundations of Computational Intelligence, Volume 1: Learning and Approximation. Volume 201 of Studies in Computational Intelligence. Springer (2009) 159–184
6. Milidiú, R.L., dos Santos, C.N., Duarte, J.C.: Portuguese corpus-based learning using etl. Journal of the Brazilian Computer Society **14**(4) (2008)
7. dos Santos, C.N., Milidiú, R.L., Rentería, R.P.: Portuguese part-of-speech tagging using entropy guided transformation learning. In: Proceedings of PROPOR 2008, Aveiro, Portugal (Setember 2008) 143–152
8. Fernandes, E.R., Pires, B.A., dos Santos, C.N., Milidiú, R.L.: Clause identification using entropy guided transformation learning. In: Proceedings of STIL09. (2009)
9. Milidiú, R.L., Crestana, C.E.M., dos Santos, C.N.: A token classification approach to dependency parsing. In: Proceedings of STIL09. (2009)
10. Fernandes, E.R., dos Santos, C.N., Milidiú, R.L.: A machine learning approach to portuguese clause identification. In: Proceedings of PROPOR 2010. (2010) 55–64
11. Fernandes, E.R., dos Santos, C.N., Milidiú, R.L.: Portuguese language processing service. In: Proceedings of WWW in Ibero-America Alternate Track of the 19th International World Wide Web Conference. (2009)
12. Breiman, L.: Bagging predictors. Machine Learning **24**(2) (1996) 123–140
13. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. **20**(8) (1998) 832–844