# Adaptive Information Extraction from Web Pages by Supervised Wrapper Induction

Rinaldo José de Lima[1], Frederico Luiz Gonçalves Freitas[1], Bernard Espinasse[2]

[1] Centro de Informática Universidade Federal de Pernambuco Recife PE Brazil,
{rjl4,fred}@cin.ufpe.br
[2] LSIS UMR CNRS 6168 Université d'Aix-Marseille Marseille France,
bernard.espinasse@lsis.org

## 1 Introduction

In this work, we are concerned with Information Extraction (EI) which comprises techniques and algorithms performing two important tasks: identifying the desired, relevant information from semi-structured or non-structured documents and storing it in appropriate formats for future use. Our focus is adaptive IE systems that can be customized for new domains through training using annotated *corpora* as input. Particularly, we look into automatic *wrapper induction* and Natural Language Processing (NLP) techniques for extraction that rest on the exploitation of structural and grammatical regularities present in documents. Wrappers are procedures to extract data from information resources. Wrapper induction is a technique that uses machine learning algorithms for automatically construct wrappers from a previously annotated corpus [4]. The wrappers we developed are based on the Boosted Wrapper Induction (BWI) algorithm [1] and integrate IE system TIES (Trainable Information Extraction System)[6], developed at ITC-irst. BWI uses the *AdaBoost* algorithm that works by continually reweighting the training examples, and using a base learner (called weak learner) to learn a new classifier repeatedly, stopping after a fixed number os iterations. The classifiers learned are then combined by weighted voting [5]. TIES incorporates the BWI algorithm and automatically induces wrappers from a set of documents annotated with XML tags that identify instances of entities to be extracted. Kauchak [3] has investigated how boosting contributes to the success of the BWI algorithm and studied its performance in the challenging direction of using it as an IE method for unstructured natural language documents. This fact motivated us to extend TIES current version to include Parts-of-Speech (POS) tagging in its preprocessing phase using a POS tagger. Moreover, we have included a module for cleaning up ill-formed tags and attributes of Web pages to produce well-formed XHTML documents which are submitted for tokenisation in TIES architecture.

## 2 Experiments and Results

The following tables show the first results obtained using the newly extended TIES system on standard tasks for adaptive IE: the CMU Seminars and Austin Jobs announcements and Call for Papers (Pascal Challenge) [2]. We measured the performance in terms of the classical measures in IE domain: *precision*, *recall*, and *F-measure*. We also conducted experiments using various combinations of features in order to systemantically examine their effects on the perfomance of the learning algorithm based on supervised classification.

Table 1: Results without POS Information

| Corpus | Precision | Recall | F1-Measure |
|---|---|---|---|
| Seminars | 0.974 | 0.953 | 0.963 |
| Jobs | 0.945 | 0.778 | 0.853 |
| CFP | 0.891 | 0.571 | 0.696 |

Table 2: Results with POS Information

| Corpus | Precision | Recall | F1-Measure |
|---|---|---|---|
| Seminars | 0.971 | 0.964 | 0.967 |
| Jobs | 0.939 | 0.780 | 0.853 |
| CFP | 0.896 | 0.591 | 0.712 |

# 3 Conclusion and Future Work

The results we obtained allow us to conclude that the newly extended TIES system, an adaptive IE system based on supervised wrapper induction is comparable with other state-of-the-art IE systems on traditional IE tasks.

Future work includes the improvement of TIES architecture by including new supervised machine learning algorithms, such as Support Vector Machines and C4.5 as learning components for new IE wrappers. Another main issue that is left for future work is related to the tokenizer and feature extraction modules in which we intend to perform the following NLP subtasks, i.e, Named Entity Recognition and Chunking Analysys (for English) and POS tagging for Portuguese and French languages.

# References


[1] Freitag, D., Kushmerick, N.: Boosted Wrapper Induction. In: Proceedings of the Seventeenth National Conference on Artificial Intelligence. AAAI-2000,(2000).

[2] Ireson, N., Ciravegna, F.: Pascal Challenge: The Evaluation of Machine Learning for Information Extraction. Machine Learning for the Semantic Web. Dagstuhl Seminar,Dagstuhl, DE (2005).

[3] Kauchak, D., Smarr, J., Elkan, C.: Sources of Success for Information Extraction Methods. Technical Report CS2002-0696. Department of Computer Science and Engineering. University of California, San Diego, January (2002).

[4] Kushmerick, N.: Wrapper Induction: Efficiency and Expressiveness. Artificial Intelligence. vol. 118, (2000).

[5] Shapire, R. E.: A Brief Introduction to Boosting. In Proceedings of the 16th International Joint Conference on Artificial Inteligence (IJCAI) (1999).

[6] TIES. Trainable Information Extraction System. http://tcc.itc.it/research/textec/tools-resources/ties.html, (2004).