Paper presented at the: *10th European Conference on Machine Learning, (ECML'98) - Text Mining Workshop*
Available at `http://www-poleia.lip6.fr/~zaragoza`

An Automatic Surface Information Extraction System Using Hierarchical IR and Stochstic IE.

Hugo Zaragoza, Patrick Gallinari.

LIP6, *Université Pierre et Marie Curie*,
4, place Jussieu F-75252 PARIS cedex 05 (F).
{Hugo.Zaragoza, Patrick.Gallinari}@lip6.fr

Abstract

We address the problem of constructing an automated information extraction system from a corpus of text. We present in this paper a novel approach to textual information extraction (IE) combining information retrieval (IR) techniques and stochastic language modelling in a hierarchical system. This permits to largely reduce the amount of data treated with complex (slow) analysis systems, it allows for a successive refinement of the features extracted, and shows better performances than non-hierarchical models. At the lowest level of the hierarchy, documents and paragraphs are successively filtered with IR techniques. At the top level, a stochastic language model extracts the most relevant phrases, according to an extraction task. The approach and preliminary results are demonstrated on a subset of the MUC-6 Scenario Templates task.

1 Introduction

Information Extraction (IE) and Information Retrieval (IR) of textual data have recently become the subject of intensive research in machine learning and artificial intelligence. There is a need to rapidly prototype systems capable of analyzing large corpus of textual information. We investigate in our present work the simultaneous application of two instances of this research, namely the *routing* problem of IR and the *message understanding* approach to IE.

Our goal in this work is to detect and extract passages containing information relevant to a message understanding task. In order to test our approach we use the MUC-6 corpus and associated Scenario Templates (ST) (Grishman, 1996). This corpus contains 200 articles from the Wall Street Journal. For each job appointment, reassignment or destitution appearing in these articles, there is an associated ST (an object which describes the event, see Fig. 1). The corpus contains 200 articles and 400 scenarios. However, 50% articles and more than 80% of paragraphs are irrelevant to the extraction task. In typical Internet applications the amount of irrelevant data is even greater.

We want to reject information which is irrelevant prior to the extraction task. The approach presented here is original in that it combines IR to IE in order to: i) eliminate as early as possible irrelevant information ii) successively refine the selected features (terms) on the relevant passages. (Fig. 1b). Specifically, we will use IR routing techniques for detection at

the document and paragraph level, and IE techniques for extraction at word level.

We will finish this section with an introductin of Information Retrieval and Extraction are . In Section 2 we describe in more detail IR in the context of our hierarchical model for the document and paragraph levels. In section Section 3 we present our Hidden Markov models for the extraction of phrases, and in Section 5 we briefly discuss research perspectives.

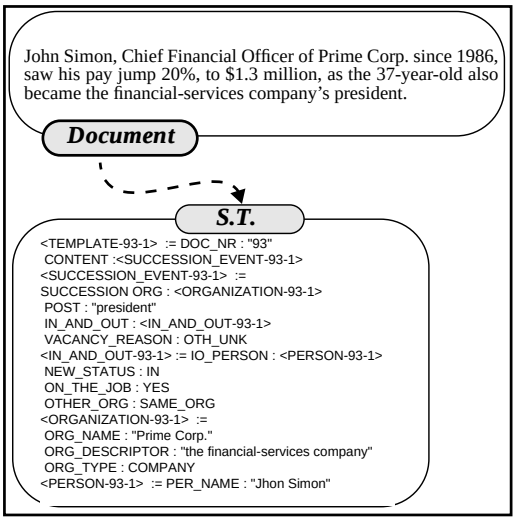

Fig. 1. a) An example from the MUC-6 Scenario Template Task. The corpus consists of Wall Street Journal Articles (*Documents*) and their corresponding filled templates (*STs*).

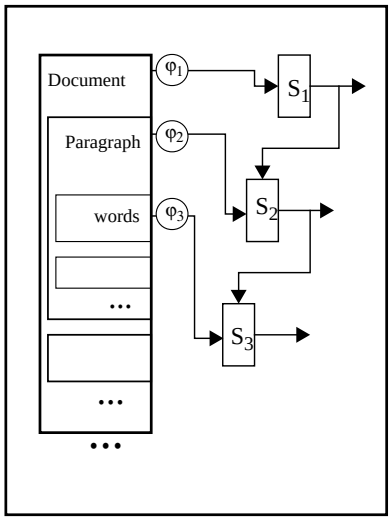

Fig. 1. b) The Hierarchical Information Extraction System. Preprocessors (φ_i) are represented by ovals and classifiers (S_i) by boxes.

Information Retrieval and Extraction

The initial goal of Information Retrieval (IR) was to rapidly index, from very large databases, subsets (possibly ranked) of documents relevant to a particular user-query (Sutton, 1989). Typical applications are library databases: a user wishes to find the most relevant documents with respect to a particular query, but IR tasks have greately diversified in the last years. We deal here only with the filtering problem of IR, namely, we will assume that user queries are defined in advance.

Information Extraction (IE) aims to (automatically) interpret the information in textual data, to parse the underlying messages; this is needed, for example, to interpret user requests, constitute databases from written documents (Lehnert, 1996). IE in the past has been applied to tasks neatly confined in a strict domain of knowledge. Systems were hand-crafted by careful analysis of the linguistic relations of words in a given context. This worked reasonably well for constrained text but was not extendable nor re-usable for «noisy» or ungrammatical text (such as speech retranscriptions, news-postings, etc.). In order to overcome these problems, IE systems recently began to integrate learning modules and ML techniques (Grishman and Sundheim, 1996).

To our knowledge there are no previous approaches combining hierarchical architectures and stochastic language modelling for IE. Hierarchical architectures in IR have been used for example by Koller and Sahami (1997), who implement two-level probabilistic decision trees for text categorization, and by Wiene, Pedersen and Weigen (1995) who take a similar approach using neural networks. The combination of IE and IR techniques is the subject of recent investigation in symbolic ML (Robertson, 1997). Stochastic language models have been used for filtering and highlighting as described in (Knaus et al., 1995).

2 Hierarchical Retrieval (Document and Paragraph levels)

Standard IR techniques treat documents as vectors in the *document-space*, where each coordinate codes the frequency of appearance of a particular *term* in the document. Terms are a subset of words which have been pre-selected by a *feature selection* method. Knowing the class of vector-coded documents, one may construct a classifier.

We consider here only two document classes: relevant (R) and irrelevant (I). A passage is relevant if there is at least one ST associated with it. We construct two standard IR modules dealing with the document and paragraph levels of representation (from now on we will denote by *passage* a document or paragraph indistinctly). The architecture of the two modules is identical, the only difference being the size of the passages and the features selected.

We use the U-measure (Andersen, 1992) for feature selection. If we call p and p' respectively the number of relevant and irrelevant passages in which the word w_i appears, q and q' the number of relevant and irrelevant passages in which it does not appear, and N the total number of passages (N = p + p' + q + q') then:

$$u_i = \varphi_U(w_i, B) = \sqrt{N} \cdot \frac{(pq' - p'q)}{\sqrt{(p + p') \cdot (p + q) \cdot (q + q') \cdot (p' + q')}} \qquad (1)$$

where w_i and u_i are a word and its U-value respectively and B is the labelled corpus. We compute the U-value of all distinct words in the corpus (after some standard pre-coding) and select as terms the 100 words with highest U-value.

The standard IR technique of *tf-idf* weighting is then used to encode passages into 100 dimensional vectors $\mathbf{x}^j = (x^j_1, x^j_2, ..., x^j_{100})^T$ where j is the passage index and:

$$x^j_i = \varphi_{passage}(t_i, P_j, B) = tf(t_i, P_j) \cdot idf(t_i, B)^2 \quad , \qquad (2)$$

tf(t_i,P_j) is the feature frequency of term t_i in passage P_j and *idf* is the inverse document frequency of term t_i in corpus B (Sutton, 1989).

Multi-layer perceptrons are then used to obtain a rough approximation of the posterior probabilities $P(R|\mathbf{x})$ and $P(I|\mathbf{x})$ of the two classes given a passage. We can then apply standard decision theory to choose the most probable class of a previously unseen passage: $class = \text{argmax}_{R,I}\{P(R|\mathbf{x}) \cdot c_{IR},\ P(I|\mathbf{x}) \cdot c_{RI}\}$, where c_{IR} and c_{RI} are the costs of misclassifcation of each class respectively.

Note that, since rejected information is lost from further computations, it is important that models reject as little as possible relevant information. Accepting irrelevant information, on the other hand, is not critical (except perhaps at the word level). For this reason, we are not interested in the classification performance but, rather, on the precision-recall performance[1], which is the classification performance as c_{IR} goes from 1 (recall=0) to 0 (recall=1).

Because the U formula depends on the frequency of appearance of each term in relevant and non relevant sections, we expect the feature selection to be more accurate at the paragraph level if we first eliminate most irrelevant documents from the base. We test this by classifying paragraphs with and without the document module. In the first case, documents are first filtered through the document classification module, and only those documents with high probability of relevance are considered for paragraph classification. In Fig. 2 we show the precision-recall curves for these two systems: we see that the hierarchical system (*Doc+Par*) is clearly better than the non-hierarchical one (*Par*) for regions of high values of recall, which are the regions in which we are interested.

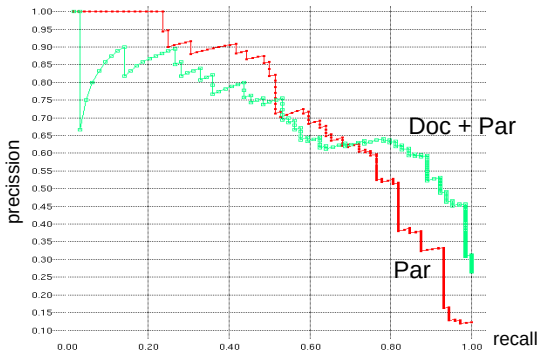

Fig. 2. Precision-recall performance for the hierarchical (Doc+Par) and non -hierarchical (Par) models.

3 The Stochastic Extraction Model (Word level)

The previous hierarchical model eliminates irrelevant documents and paragraphs. This reduces the mass of documents analyzed to a small number of paragraphs, from which we then wish to extract the relevant information. The word level however is intrinsically different from phrase, paragraph or document levels in that context needs to be taken into account. We deal with this level of representation with Hidden Markov Models (HMMs).

HMMs allows us to construct *passage models* which are simplified descriptions of the probability distribution of words in passages. We consider words as the observable (output) symbols of an unknown (hidden) sequence of states. The probability distribution of terms in the states can be obtained by counting on a labelled set. The parameters of the model (state tran-

1. Let *n* be the number of documents chosen as relevant by our classifier, *R* the total number of relevant passages and *r* the number of relevant passages classified as relevant. Then *precision* = *r*/*n*, and *recall* = *r*/*R*.

sition probabilities as well as term emission probabilities) can then be learned with standard HMM techniques (Rabiner, 1993)[1].

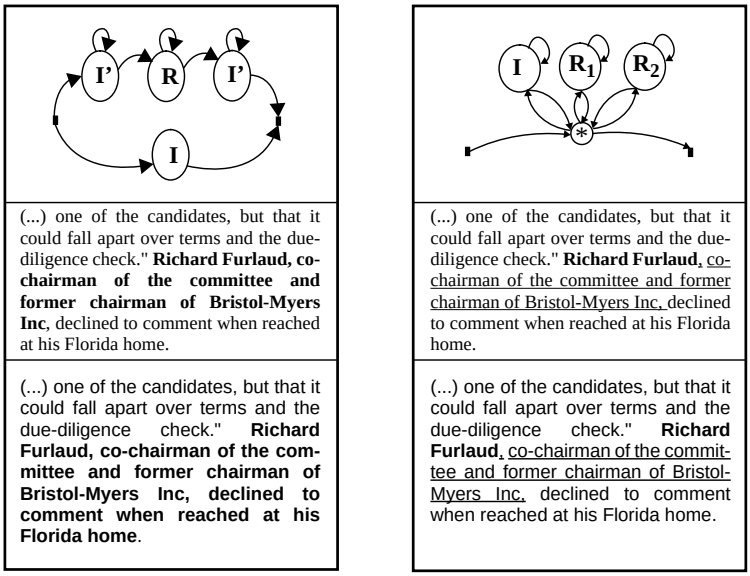

Fig. 3. 4) Two HMMs for language modelling (top) and an instance of the extraction task. A section of a relevant paragraph and its desired tagging is shown in the middle and the obtained with the model is shown at the bottom. On the left, bold indicates relevant (R) text. On the right bold and underline indicate *PERSON* (R1) and *POSITION* (R2) states respectively.

The number of distinct words is however too large. We use the U-measure (eq. 1) to map words into a one-dimensional space (in previous work Mittendorf and Schauble (1994) proposed a similar model for IR and filtering using the *tf-idf* mapping). With this measure we may map sequences of words (or passages) (w_1,w_2,...w_T) into sequences of real values (u_1, u_2, ..., u_T), on which continuous HMMs operate.

As a first example, we propose a simple passage model (Fig. 3, top left) similar to the model proposed in (Mittendorf and Schauble, 1994). In this simple model, sequences of words must follow one of two paths when «traversing» this HMM: either all words are considered irrelevant, or a sequence of consecutive irrelevant-relevant-irrelevant sub-passages is found. In the first case, the passage is considered irrelevant, in the second case the part of the text matching the relevant state is extracted. Bellow the HMM, we see the original text (Fig. 3, middle left) with the passage containing the relevant information in bold style. Following this text (Fig. 3, bottom left) we see the most probable path (with words matching

1. Labels where assigned to words of the relevant MUC-6 documents as follows: a paragraph containing strings matching at least two slots of an ST where considered relevant; phrases from relevant paragraphs containing at least one string matching a slot of the associated ST where labelled with the slot's name.

the R state in bold) of the HMM after training.

The previous example only extracted relevant passages, without distinctions to the type of information to be found.This could be used to skim large databases and select only relevant information, which would then be treated by a more complex IE system. As a more complex example of information extraction we introduce a model (Fig. 3, top right) that tags automatically portions of text. Three tags are used: irrelevant (I), *PERSON* (R1) and *POSITION* (R2). The last two tags correspond to slots of the MUC-6 ST objects (Fig. 1a). We are interested in the segmentation produced by the HMM, specifically in the mapping between words and states. In the example (Fig. 3 right), bold and underlined represent the *PERSON* and *POSITION* tags respectively. The sequence of states obtained by the model is I-R1-R2-I, R1 coinciding with "Richard Furlaud" and R2 with "co-chairman (...) Bristol-Mayers Inc." which are the desired PERSON and POSITION slots). The *star* state assures that transitions happen only at the beginning of phrases (defined as sequences of words separated by punctuation marks). While this is not necessary, it speeds up the search process and produces more coherent sequences. For the moment we have observed good behavior of our models, but we are still working on the evaluation of our models.

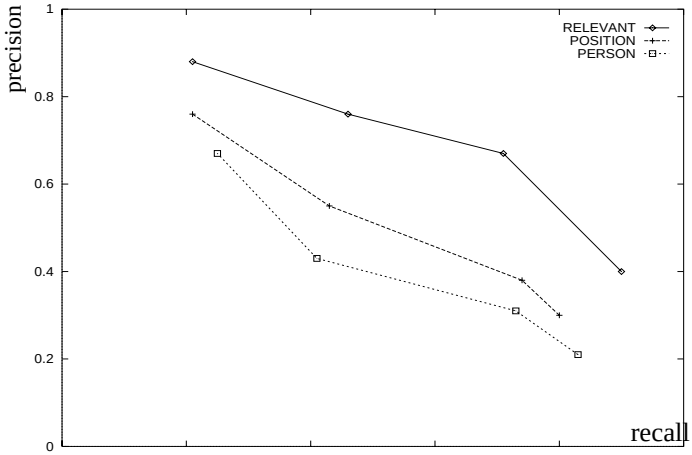

Fig. 4. Preliminary Performances of the stochastic IE module.

4 Results

The evaluation of IR and IE systems is a difficult problem. As a first evaluation of our extraction module, we use the measures of precision and recall as follows. Let h be the number of words correctly labelled by the model, N the true total number of words of a given concept, and H the total number of words labelled by the model as being of that concept, then for this concept, precision = h/H and recall = h/N. In order to obtain different precision-recall points we fix different values for the state transition probabilities (increasing or decreasing the probability of staying in the concept state). In Fig. 4 we present the precision-recall curves for the first and second models discussed (Figure 4). For the second model, we present two precision-recall curves: one for the POSITION field and one for the PERSON field. Evaluation is still quite preliminary, but it is encouraging. We see that simple models are capable of simple surface IE.

5 Perspectives

We have outlined the most important aspects of our work on a hierarchical approach to automatic information retrieval and extraction. IR techniques are used to eliminate documents which are irrelevant to the extraction task, then paragraphs. Finally, an stochastic model extract (and labels) the sequences of words which contain the information to be extracted.

We use our approach to skim, condense and pre-code textual data so that more complex conceptual taggers (such as those we developed for constrained information retrieval tasks (Stévenin-Barbier, 1996)) can be re-used. Furthermore, we wish to extend the stochastic models to make a richer use of context. Parallel to this work, we are investigating approaches to learn simultaneously the different models of our hierarchy. In this context, we study coding schemes for adaptive feature selection. Finally, rigorous evaluation of our models is underway.

References

Andersen E. (1992) *The Statistical Analysis of Categorical Data.* Springer, Berlin.

Grishman R. (1996) Design of the MUC-6 Evaluation. *Proceedings of the Sixth Message Understanding Conference (MUC-6)* (Columbia 1995), 13-33. Morgan Kauffman, San Francisco, CA.

Grishman R. et Sundheim B. (1996) Message Understanding Conference - 6 : A Brief History. *COLING 96, 16th International Conference on Computational Linguistics*. (Copenhagen 1996) vol. 1, 466-471.

Harman D. (1996) Overview of the Fourth Text REtrieval Conference (TREC-4) *Proc. of the 4th Text Retrieval Conference* (Washington 1996).

Knaus D., Mittendorf E., Schauble P. and Sheridan P. (1995) Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System. *Proc. of the 4th Text Retrieval Conference TREC-4.*

Koller D., Sahami M.(1996) *Hierarchically classifying documents using very few words*.

Lehnert, W. (1996) *Information Extraction.* (electronic publication: http ://www-nlp.cs.umass.edu/ nlpgroup/nlpie.html)

Mittendorf E., et Schauble P. (1994) Document and Passage Retrieval Based on Hidden Markov Models. *ACM SIGIR*'94, 318-327.

Rabiner L., Juang B.H. (1993) *Fundamentals of Speech Recognition*, Prentice Hall Signal Processing Series, 1993.

Robertson M.A. (1997) On the Marriage of Information Retrieval and Information Extraction. *19th Annual BCS-IRSG Colloquium on IR Research* (Aberdeen 1997) 60-69 Furner J. Harper D. (eds.)

Stévenin-Barbier A. et Gallinari P. (1997) Semantic anticipation for understanding using neural networks, *PACES / SPICIS* (Singapore 1997).

Sutton G. (1989) *Automatic Text Processing : the Transformation, Analysis and Retrieval of Information by Computer.* Addison Wesley 1989.

Wiener, Pedersen (1995). A Neural Network Approach to Topic Spotting. In *Proc. of the Foruth Anual Symp. on Doc. Analysis and Information Retrieval (SDAIR'95)*, 317-321.