In: *Proceedings of CoNLL-2000 and LLL-2000*, pages 194–198, Lisbon, Portugal, 2000.

Learning from Parsed Sentences with INTHELEX

F. Esposito and **S. Ferilli** and **N. Fanizzi** and **G. Semeraro**
Dipartimento di Informatica
Università di Bari
via E. Orabona, 4 - 70126 Bari - Italia
{esposito, ferilli, fanizzi, semeraro}@di.uniba.it

Abstract

In the context of *language learning*, we address a logical approach to information extraction. The system INTHELEX, used to carry out this task, requires a logic representation of sentences to run the learning algorithm. Hence, the need for parsers to produce structured representations from raw text. This led us to develop a prototypical Italian language parser, as a preprocessor in order to obtain the structured representation of sentences required for the symbolic learner to work. A preliminary experimentation proved that the logic approach to learning from language is able to capture the semantics underlying the kind of sentences that were processed, even if a comparison with classical methods as regards efficiency has still to be done.

1 Introduction

Language learning has gained growing attention in the last years. Statistical approaches, so far extensively used — see (Saitta and Neri, 1997) for an overview of the research in this area —, have severe limitations, whereas the flexibility and expressivity of logical representations make them highly suitable for natural language analysis (Cussens, 1999). Indeed, logical approaches may have a relevant impact at the level of semantic interpretation, where a logical representation of the meaning of a sentence is important and useful (Mooney, 1999).

Logical approaches have been already employed in Text Categorization and/or Information Extraction. Yet they try to use an expressive representation language such as first-order logic to define simple properties about textual sources, regarded, for instance, as bags of words (Junker et al., 1999) or as semi-structured texts (Freitag, 2000). These properties are often loosely related with the grammar of the underlying language, often relying on extra-grammatical features (Cohen, 1996). We intend to exploit a logic representation for exploiting the grammatical structure of texts, as it could be detected using a proper parser. Indeed, a more knowledge intensive technique is likely to perform better applied on the tasks mentioned above.

When no background knowledge about the language structure is assumed to be available, one of the fundamental problems with the adoption of logic learning techniques is that a structured representation of sentences is required on which the learning algorithm can be run. Thus, the need arises for parsers that are able to discover such a structure starting from raw, unstructured text. Research in this field has produced a variety of tools and techniques for English, that cannot be applied to other languages, such as Italian, because of the different, and sometimes much more complex, grammatical structure. Such considerations led us to develop a prototypical Italian language parser, that could serve as a pre-processor of texts in order to obtain the structured representation of sentences that is needed for the symbolic learner to work. It is fundamental to note that the focus of this paper is not the parser, that does not adopt sophisticated NLP techniques. The aim here is investigating the feasibility of learning semantic definitions for some kinds of sentences. Even more so, the availability of a professional parser will further enhance the performance of the whole process.

Further problems in applying relational learning to language are due to the intrinsic computational complexity of these methods, as a drawback of the expressive power gained through relations. Moreover, another weakness of our approach could be the dependence on the quality of the data coming from the preprocessing step: it is possible that noise coming from wrongly parsed sentences be present, thus having a negative influence towards the model to be induced.

After briefly presenting in Section 2 the parser performance, in order to establish the degree of reliability of the data on which the learning step is performed, Section 3 shows the results of applying the first-order learning system INTHELEX (Esposito et al., 2000) for the inference of some simple events related to foreign commerce. Lastly, Section 4 draws some preliminary conclusions on this research and outlines future work issues to be addressed.

2 A Stratified Parser for Italian Language

This section presents a parser for the Italian language, based on context-free grammars and designed to manage texts having a simple and standard phrase structure (e.g., foreign commerce texts as opposed to poetry texts). It is composed by 12 parsing levels and 106 production rules, and uses the longest-match technique, which complies with the typical ambiguity of Italian language. Syntactic lookahead is used to overcome ambiguity and to prevent the parsing from stopping in case of grammatically wrong input.

The text is segmented in progressively larger syntactic constructs. Subject, main verb, direct or indirect object and clauses referring to them are identified. Nested syntactic constructs at the same abstraction level (e.g., expressions including a sentence in parentheses) are supported.

Plain text documents are provided to a lexical analyzer and a noun-recognizer (XEROX MULTEXT), whose output is the document text tagged with parts of speech to be fed to the parser. Since Italian grammar is very different from the English one, some terms do not have an English equivalent and, hence, cannot be translated.

The parser was validated on a set of 72 sentences drawn from a corpus of articles on foreign commerce available on the Internet, and the results obtained were evaluated with respect to *precision*, *recall* (reported in Table 1) and two measures about error ratio:
Error1 = # errors/# total constituents extracted
Error2 = # errors/# total correct constituents expected
(see Table 2).

Parsing Phase	**Precision**	**Recall**
Noun Groups	0.984	0.992
1st level NPs	0.994	0.994
2nd level NPs	0.983	0.983
PPs	0.951	0.951
Clauses	0.840	0.840
Refined clauses	0.913	0.913
Sentences	0.736	0.736

Table 1: Summary of parser validation results (Precision/Recall)

Parsing Phase	**Error1**	**Error2**
Noun Groups	0.787%	0.793%
1st level NPs	0.596%	0.596%
2nd level NPs	1.666%	1.666%
PPs	4.918%	4.918%
Clauses	15.941%	15.941%
Refined clauses	8.695%	8.695%
Sentences	26.384%	26.384%

Table 2: Summary of parser validation results (Error1/Error2)

3 Information extraction

The grammar above was used to parse Italian texts downloaded from the Internet, and concerning foreign commerce. Through such preprocessing, the aim was to obtain some structure for those texts that could then be translated in the input language of the learning system INTHELEX (Esposito et al., 2000) in order to make it learn simple events concerning that domain.

INTHELEX (INcremental THEory Learner from EXamples) is a fully incremental, multi-conceptual closed loop learning system for the induction of hierarchical theories from examples. In detail, full incrementality avoids the need of a previously generated version of the theory to be available, so that learning can start from an empty theory and from the first example; multi-conceptual means that it can learn simultaneously various concepts, possibly related to each other; a closed loop system is a system in which the learned theory is checked to be valid on any new example available, and in case of failure a revision process is activated on it, in order to restore the completeness and consistency properties.

Incremental learning is necessary when either incomplete information is available at the time of initial theory generation, or the nature of the concepts evolves dynamically. The latter situation is the most difficult to handle since time evolution needs to be considered. In any case, it is useful to consider learning as a closed loop process, where feedback on performance is used to activate the theory revision phase.

INTHELEX learns theories, from positive and negative examples described in the same language. It adopts a full memory storage strategy — i.e., it retains all the available examples, thus the learned theories are guaranteed to be valid on the whole set of known examples.

In the formal representation of texts, we used the following descriptors:

- `sent(e1,e2)` e2 is a sentence from e1
- `subj(e1,e2)` e2 is the subject of e1
- `obj(e1,e2)` e2 is the (direct) object of e1
- `indirect_obj(e1,e2)` e2 is an indirect object of e1
- `rel_subj(e1,e2)` e2 is a clause related to the subject e1
- `rel_obj(e1,e2)` e2 is a clause related to the object e1
- `verb(e1,e2)` e2 is the verb of e1
- `lemma(e2)` word e2 has lemma *lemma*
- `infinite(e2)` verb e2 is in an infinite mood
- `finite(e2)` verb e2 is in a finite mood
- `affirmative(e2)` verb e2 is in an affirmative mood
- `negative(e2)` verb e2 is in a negative mood
- `np(e1,e2)` e2 is a 2nd level NP of e1
- `pp(e1,e2)` e2 is a PP of e1

where *lemma* is a meta-predicate. This allows the system to exploit information about word lemmas in generalizations/specializations, and in the recognition of higher level concepts of which *lemma* is an instance.

Thus, the following Horn clause is an instance of an example:

```
imports(example) ←
          sent(example,e1),
          subj(example,e2),
          np(e2,e3),
          impresa(e3),
          rel_subj(e1,e4),
          verb(e4,e5),
          specializzare(e5),
          infinite(e5),
          affirmative(e5),
          pp(e4,e6),
          distrubuzione(e6),
          componente(e6),
          verb(e1,e7),
          interessare(e7),
          finite(e7),
          affirmative(e7),
          indirect_obj(e1,e8),
          pp(e8,e9),
          importazione(e9),
          macchina(e9),
          produzione(e9),
          ombrello(e9).
```

A first experiment aimed at learning the concept of specialization (of someone in some field). The system was run on 40 examples, 24 positive and 16 negative. The resulting theory was made up by 5 clauses, some of which differ just in one literal (e.g., the lemma of the word in the subject). By exploiting the background knowledge that terms 'impresa', 'società', 'ditta' and 'agenzia' are all instances of the concept 'persona giuridica', i.e. clauses:

```
persona_giuridica(X) ←
          ditta(X).

persona_giuridica(X) ←
          societa(X).

persona_giuridica(X) ←
          impresa(X).

persona_giuridica(X) ←
          agenzia(X).
```

the theory becomes more compact, yielding the following rules:

```
specialization(A) ←
          sent(A,B),
          subj(B,C), np(C,D),
          persona_giuridica(D),
          verb(B,E),
          specializzare(E),
          finite(E),
          affirmative(E),
          indirect_obj(B,F),
          pp(F,_).

specialization(A) ←
          sent(A,B),
          subj(B,C), np(C,D),
          persona_giuridica(D),
          rel_subj(B,E),
          verb(E,F),
          specializzare(F),
          affirmative(F),
          pp(E,_), verb(B,_).
```

Another experiment aimed at learning the concept of "imports". INTHELEX was run starting from the empty theory, and was fed with a total of 67 examples (39 positive and 28 negative). It should be noted that not all positive examples explicitly use verb 'importare' (to import): e.g., in the sentence "Società belga, specializzata nella lavorazione del legno, cerca fornitori di legname" the imports event is characterized by the noun 'società' (society) as the sentence subject, by the verb 'cercare' (to look for) and by the object including the noun 'fornitore' (provider). We obtained the following results (in which the above background knowledge was used to compress more rules into one, too):

```
imports(A) ← sent(A,B),
          subj(B,C), np(C,D),
          persona_giuridica(D),
          verb(B,E),
          cercare(E),
          finite(E),
          affirmative(E),
          obj(B,F), np(F,G),
          fornitore(G).

imports(A) ← sent(A,B),
          subj(B,C), np(C,D),
          persona_giuridica(D),
          societa(D),
          verb(B,E),
          cercare(E),
          finite(E),
          affirmative(E),
          obj(B,F), np(F,G),
          distributore(G).

imports(A) ← sent(A,B),
          subj(B,C), np(C,D),
          persona_giuridica(D),
          verb(B,E),
          interessare(E),
          finite(E),
          affirmative(E),
          indirect_obj(B,F),
          pp(F,G),
          importazione(G).

imports(A) ← sent(A,B),
          subj(B,C), np(C,D),
          persona_giuridica(D),
          verb(B,E),
          acquistare(E),
          finite(E),
          affirmative(E).

imports(A) ← sent(A,B),
          subj(B,C), np(C,D),
          persona_giuridica(D),
          impresa(D),
          verb(B,E),
          importare(E),
          finite(E),
          affirmative(E).
```

For instance, the third clause means: "Text A deals with imports if it contains a sentence with a subject composed by a NP containing a *persona giuridica*, the verb of the main sentence is *interessare* (to interest) in finite affirmative mood, and the indirect object consists of a PP containing the word *importazione*". Note that, by exploiting a background knowledge that represents a more complex ontology than the current one, it would be possible to further merge conceptual descriptors and, as a consequence, clauses in the theory. For example, 'fornitore' (provider) and 'distributore' (distributor) could be recognized as instances of a common higher level concept; the same applies to 'acquistare' (to buy) and 'importare' (to import).

4 Conclusions & Future Work

We have addressed the problem of learning logic theories for information extraction so to benefit by the semantic interpretation provided by a logical approach. This has required structured sentences in a logic representation on which to run our learning algorithms. Hence, we needed a parser to produce structured representations from raw unstructured text. Though many techniques have been developed for English, they cannot be applied to other languages, such as Italian, because of the different grammatical structure. This has led us to develop a prototypical Italian language parser, as a preprocessor in order to obtain the structured representation of sentences needed for the symbolic learner to work.

Future work will concern a more extensive experimentation, an empirical evaluation of our approach, and application of the same kind of experiments on English parsed texts. If good results will be obtained, it is possible thinking to carry out experiments that take advantage also from the structure of semi-structured documents. Indeed, we are involved in the project CDL (Esposito et al., 1998; Costabile et al., 1999), that could profit by this kind of techniques as regard semantic indexing of the stored documents (cf. (Chanod, 1999)).

References

J.-P. Chanod. 1999. Natural language processing and digital libraries. In M.T. Pazienza, editor, *Information Extraction*, volume 1714 of *Lecture Notes in Artificial Intelligence Tutorial*, pages 17–31. Springer.

W. Cohen. 1996. Learning to classify english text with ilp methods. In Luc de Raedt, editor, *Advances in Inductive Logic Programming*, pages 124–143. IOS Press, Amsterdam, NL.

M.F. Costabile, F. Esposito, G. Semeraro, and N. Fanizzi. 1999. An adaptive visual environment for digital libraries. *International Journal of Digital Libraries*, 2:124–143.

J. Cussens, editor. 1999. *Learning Language in Logic*. Workshop on Learning Language in Logic, Workshop Notes.

F. Esposito, D. Malerba, G. Semeraro, N. Fanizzi, and S. Ferilli. 1998. Adding machine learning and knowledge intensive techniques to a digital library service. *International Journal of Digital Libraries*, 2(1):3–19.

F. Esposito, G. Semeraro, N. Fanizzi, and S. Ferilli. 2000. Multistrategy Theory Revision: Induction and abduction in INTHELEX. *Machine Learning*, 38(1/2):133–156.

D. Freitag. 2000. Machine learning for information extraction in informal domains. *Machine Learning*, 39(2/3):169–202.

M. Junker, M. Sintek, and M. Rinck. 1999. Learning for text categorization and information extraction with ILP. In James Cussens, editor, *Proceedings of the First International Workshop on Learning Language in Logic - LLL99.*

R. Mooney. 1999. Learning for semantic interpretation: Scaling up without dumbing down. In Cussens (Cussens, 1999), pages 7–15. Workshop on Learning Language in Logic, Workshop Notes.

L. Saitta and F. Neri. 1997. Machine learning for information extraction. In M.T. Pazienza, editor, *Information Extraction*, volume 1299 of *Lecture Notes in Artificial Intelligence Tutorial*, pages 171–191. Springer.