Hybrid Constraints for Robust Parsing: First Experiments and Evaluation

Roberto Bartolini[1], Alessandro Lenci[2], Simonetta Montemagni[1], Vito Pirrelli[1]

Istituto di Linguistica Computazionale – CNR[1]
Area della Ricerca, via G. Moruzzi 1, 56100 Pisa, Italy
{roberto.bartolini, simonetta.montemagni, vito.pirrelli}@ilc.cnr.it

Università di Pisa, Dipartimento di Linguistica[2]
via Santa Maria, 36, 56100 Pisa, Italy
alessandro.lenci@ilc.cnr.it

Abstract

In this paper we present IDEAL+, a parsing architecture for Italian, which pursues the goal of pairing robustness with deep linguistic analysis by extending a shallow processing kernel with a pool of hybrid constraints for the incremental identification of grammatical relations. The parsing output takes the form of dependency structures representing the full range of instantiated functional relations (e.g. subject, object, modifier, complement, etc.). The paper focuses on nature and interaction of the battery of hybrid constraints and evaluates their joint impact against a gold standard of more than 700 manually annotated sentences.

Introduction

In the literature, a wide consensus has accrued on the need to overcome the persistent gap between shallow and deep language processing. While shallow parsing (*e.g.* chunking) has proved to provide useful albeit under-specified linguistic information for a number of non-trivial NLP tasks, important application-driven scenarios call for a deeper grasp of text semantics, which, in turn, presupposes a more complete description of syntactic structures. When key evaluation parameters such as coverage and robustness (in the face of non standard input) are taken into account, deep parsing architectures based on formal lexicalized grammars (e.g. HPSG, LFG) appear to be comparatively too brittle and poorly scalable to compete with shallow parsing technologies. This seems to provide *prima facie* support to the traditional view that the role played by lexical knowledge in driving syntactic analysis creates a major divide between deep and shallow parsing. If shortage of suitable information about word syntactic and semantic selective properties is, recognizably, one of the major causes for lack of coverage (and robustness) in deep parsing systems, then, so the argument goes, bigger and more complete lexicons should lead to better and deeper parsing systems. Another face of the same argument is that "plugging" a sufficiently big subcategorization lexicon into a shallow parser is a *sine qua non* of deeper levels of analysis. However true in its general outline, this argument is not a straightforward recipe for better parsing systems. Lexical information interacts with context in complex and often indirect ways. Using lexical information on the syntactic behaviour of a word irrespective of its actual use in context may in fact lead to parsing errors (Bartolini *et al.*, 2002). More complex architectures are in need which prove to be able to circumvent the traditional grammar vs. lexicon divide by promoting better integration of available lexical knowledge.

In this paper we present a battery of experiments carried out on IDEAL+, an architecture for Italian parsing, which pursues the goal of pairing robustness with deep linguistic analysis by extending a shallow processing kernel with a pool of hybrid word-based constraints jointly leading to the incremental identification of grammatical relations in context. The purpose of the reported experiments is to investigate the interplay of different types of constraints with a view to evaluating the impact of various types of lexical information in parsing performance.

The parsing system

The general architecture of IDEAL+ adheres to the key principles of *modularity* and *incrementality* (Basili and Zanzotto, 2002) by implementing a three-stage pipeline of processing modules that identify syntactic structures by progressively reducing under-specification in syntactic representations. The parsing output takes the form of dependency structures representing the full range of functional relations (e.g. subject, object, modifier, complement, etc.) within sentences.The three modules are:

1. **chunking** - a previously morphosyntactically analysed text is tagged and segmented into an unstructured sequence of text segments. Chunks represent maximally underspecified syntactic units, with a shallow internal structure and inter-chunk dependencies left unidentified;
2. **dependency analysis** – starting from chunk sequences, functional dependencies are progressively assigned to word pairs. At this stage, the output is still left widely underspecified, and functional dependencies can be ambiguously assigned to lexical heads (e.g. subject-object ambiguities are left unresolved);
3. **constraints application** – a pool of constraints is applied to the functional dependencies assigned at the previous stage to reduce ambiguities and eventually further specify grammatical relations.

Modules (1) and (2) are implemented through batteries of finite state automata that do not have access to lexical information. Module (3) includes subcategorization constraints and order preference constraints. The constraint module is currently being extended with a third layer of distributional semantic constraints, which however will not be discussed in the present experiments.

Experimental settings

Objectives

Our purpose is to investigate the contribution of different types of lexical and non-lexical constraints to the identification of grammatical dependencies. The constraints used in IDEAL+ are "hybrid" under two respects. Firstly, they use different types of linguistic information as clues to proper grammatical relation identification. Secondly they are partly data-driven and partly the results of a process of manual encoding.

In the present paper, we focus on the subtask of subject-object disambiguation in Italian. The relevance of this choice is motivated by the notorious complexity of this task in parsing Italian (Montemagni, 1995), where relatively free constituent-order allows for subjects to normally occur in postverbal positions. It is important to appreciate that only in some cases surface clues such as case-marking (limited to pronominal heads only) and verb agreement are helpful in tackling the resulting ambiguity on a purely morphosyntactic basis. Moreover, subject-object disambiguation is a typical task in which lexical subcategorization properties (such as transitivity information) are expected to play a key role.

The test corpus (TC)

The test corpus contains a selection of sentences extracted from the balanced partition of the Italian Syntactic Semantic Treebank (ISST, Montemagni *et al.*, 2003), including articles from contemporary Italian newspapers and periodicals covering a high variety of topics (politics, economy, culture, science, health, sport, leisure, etc.). TC consists of 23,919 word tokens, corresponding to 721 sentences.

The experiments

We have performed three experiments, in which different configurations of IDEAL+ have been run on TC. All configurations share the same modules (1) and (2), while differing in the constraints they use to carry out subject-object disambiguation. By relying on structural information only, modules (1) and (2) produce a set of syntactically ambiguous *candidate test pairs* (TPs), k=<v, n.*pos*>, where v_k is the verbal head and n_k is the noun dependent in the pair k. The *pos* attribute marks the position of n relative to v: pos_k = *pre*, if n_k is preverbal, and pos_k= *post*, if n_k is postverbal. In every pair $k \in TP$ generated by the system, the dependency relation linking n and v is left ambiguous between subject (*subj*) and direct object (*obj*). For each pair $k \in TP$, the constraint module (3) assigns a score σ to the interpretation of k as an instance of *subj* and to the interpretation of k as an instance of *obj*. The best scores are then used for a comparative evaluation of the different configurations wrt their ability to disambiguate subjects and objects. It is important to remark here that each disambiguation decision is taken by the system in a strictly *local* fashion. In other terms, the relation assigned to a given k does not depend on the decisions taken by the system wrt other test pairs. This surely represents an oversimplification, as the realization of a verb argument is typically not independent of the realization of another argument of the same verb. The configurations of constraints used in the three experiments are:

A – Global order constraint

This is the simplest configuration, and also acts as an experimental baseline. Here TPs are disambiguated only on the basis of a global non-lexical heuristic that assumes SVO as the unmarked order in Italian. In this case, IDEAL+ assigns to each $k \in TP$ a score σ, such that $\sigma_{k,subj}$ = 50 and $\sigma_{k,obj}$ = 40, if n_k occurs preverbally, and, conversely, $\sigma_{k,subj}$ = 40 and $\sigma_{k,obj}$ = 50 if n_k occurs postverbally. Consequently, the object interpretation of preverbal nouns is always disfavoured, consistently with the fact that this typically represents a highly marked order in Italian.

B – Global order constraint + categorical lexicon look-up

TPs are disambiguated by jointly using the global order constraint of configuration A, augmented with lexical subcategorization information. For this purpose, IDEAL+ includes a syntactic lexicon of ~26,400 subcategorization frames for nouns, verbs and adjectives derived from the Italian PAROLE syntactic lexicon. It is worth emphasizing that in this configuration subcategorization information is used in a strictly *categorical* way, in the sense of Manning (2003). That is to say, no information concerning the probability distribution of frames given a verb is available in the lexicon. As a consequence, a pair k will receive an object reading if v_k is found in the syntactic lexicon with a transitive frame. Lexical information is however combined with the global position constraint in such a way that lexically selected objects are ranked higher when they occur postverbally (if pos_k = *pre*, $\sigma_{k,obj}$ = 45; if pos_k = *post*, $\sigma_{k,obj}$ = 60). On the other hand, if v_k has an intransitive frame, $\sigma_{k,subj}$ = 60, both with preverbal and postverbal n_k.

C – Probabilistic lexical constraint

Subcategorization and word-order information are modeled in terms of probabilistic constraints. As a major difference from configurations A and B, now lexicalized argument order preferences are used instead of a global word-order heuristic. In fact, both linguistic theory and corpus data confirm that some verbs strongly prefer to realize their subjects in postverbal position. Thus, it seems plausible to assume that the lexical information relevant for subject-object disambiguation is not limited to the number and type of slots selected by a verb head, but it also includes information on the preferred position in which these arguments are realized. Therefore, although other factors are surely relevant in deciding the subject position (e.g. structural "heaviness" of the noun phrase, definiteness, information structure, etc.), a lexical word-order constraint is expected to be more reliable than a global one.

In principle, we would like our language model to express lexical constraints as rewrite probabilities. We should then be able to calculate the conditional probability of having one particular frame argument realized in context, *given* the probability that its frame is lexically selected. In turn, this requires preliminary identification of all occurrences of a frame in a training corpus for each verb of interest. This may not be an easy task. Among other things, we should always be in a position to classify – say - an SV pattern in context as either the instantiation of an intransitive frame, or that of a transitive frame with an implied direct object.

Constraints configurations	All TPs				"Gold" TPs			
	answers	prec	recall	f-score	answers	prec	recall	f-score
Configuration A	845	68.52%	100.00%	81.32%	660	87.73%	100.00%	93.46%
Configuration B	749	72.36%	88.64%	79.68%	576	94.10%	87.27%	90.56%
Configuration C	834	72.90%	98.70%	83.86%	659	92.26%	99.85%	95.90%
Configuration C with T≥1.5	801	74.16%	94.79%	83.22%	633	93.84%	95.91%	94.86%
Configuration C with T≥2	759	76.02%	89.82%	82.35%	607	95.06%	91.97%	93.49%

Table 1: Disambiguation results of Module 3

Most of the times, (manually) annotated tree banks do not provide direct evidence of this kind. If we have to rely on instantiated arguments only, as opposed to fully specified argument frames, we must then be ready to approximate frame probabilities with (conditionally independent) frame slot probabilities.
For what we said so far, we model the probability of having a direct object *n* selected in a certain position *pos* relative to its verb head *v*, as the product

$p(n, \text{pos}, obj|v) = p(tr|v)*p(n, pos, obj|tr)$,

where $p(tr|v)$ is a measure of the degree of transitivity of *v*, namely the relative number of times *v* selects an *explicit* direct object (*i.e.* an overtly realized nominal head), and $p(n, pos, obj|tr)$ is the conditional probability for a direct object *n* to be realized at *pos* every time *v* is used transitively. By the same token, we can define the probability $p(n, \text{pos}, subj|v)$ of having an overt subject at *pos* in a *v* transitive use, as follows:

$p(n, \text{pos}, subj|v) = p(tr|v)*p(n, pos, subj|tr)$.

Hence, we can meaningfully compare $p(n, \text{pos}, obj|tr)$ and $p(n, \text{pos}, subj|tr)$ to assess what is the most probable syntactic reading of a nominal in a *transitive* construction, given its position (*pos*) relative to the verb head.
In fact, in many cases, we are interested in finding the most probable syntactic reading of *n* in a construction with *one* morphosyntactically overt nominal only. Being Italian a pro-drop language, it is possible for a verb-noun sequence to be interpreted as either VO (with a pro subject) or VS (with a possibly implied object). This requires modification of

$p(n, \text{pos}, subj|v) = p(o\_subj|v)*p(n, pos, subj|\ o\_subj)$,

where $p(o\_subj|v)$ represents the probability that *v* selects an overtly realised subject in either a transitive or intransitive construction.
It is worth reminding that in Italian subject and object can also be realized both pre-verbally (SOV, OSV) or both post-verbally (VOS, VSO). For the present purposes, however, subject/object disambiguation is carried out *locally*, on the basis of information about one TP only at a time. This means that, if two nominal heads take the same position relative to the verb, they are *both* assigned the same function. Although this is a clear oversimplification, it has a minor impact on our overall results as the configurations above are extremely rare. In the whole ISST both subject and object are realized post-verbally only in 0.3% of the cases, whereas the percentage of SOV/OSV patterns is even more negligible, i.e. 0.02%.
Lexical constraints are calculated through an add-one smoothed maximum likelihood estimation of the probabilities above (Jurafsky and Martin, 2000). As a training corpus, we use the whole ISST, which includes 21,950 tokens with 2,003 verb types, 17,278 subject relations and 8,304 object relations.

Evaluation results

The three configurations of Module 3 were tested on the TPs identified in TC by IDEAL+ Modules (1) and (2). We excluded from our testbed pairs with relative and clitic pronouns (whose distribution is not lexically-governed), controlled subjects of infinitive clauses and coordinated subjects and objects. Table 1 summarises subject/object disambiguation results with different constraints configurations (namely A, B and C) against a) all TPs identified in TC (845) and b) "gold" TPs only (660), i.e. TPs attested in the reference manual annotation. By using b), we abstract away from errors originating at previous parsing stages, thus focusing on the disambiguation task only.
For each set of TPs, precision, recall and f-score figures are given for the different configurations. Precision is defined as the ratio of correctly disambiguated dependency relations over all relations disambiguated by Module 3 (prec = correctly disambiguated relations / total number of disambiguated relations); recall refers to the ratio of correctly disambiguated dependency relations over all TPs (recall = correctly disambiguated relations / |TPs|). Finally, the overall disambiguation performance of Module 3 is described in terms of the f-score, computed as follows: 2 * prec * recall / prec + recall.
As to probabilistic lexical constraints (configuration C), results are given by fixing a threshold T on the ratio between $\sigma_{k,best\_rel}$, the score assigned to the preferred interpretation, and $\sigma_{k,other\_rel}$ the score assigned to the disfavoured candidate. The assumption is that when scores assigned to competing interpretations are very close then other constraints should be resorted to for a more reliable disambiguation (see discussion below).
Results reported in Table 1 for all TPs show that precision gradually increases, from 68.52% (configuration A) to 72.36% (configuration B), up to 72.90%-76.02% (configuration C). By focussing on precision results only, a significant improvement can be observed when categorical lexical information is also taken into account (as in configuration B). Actually, this improvement is counterbalanced by a drastic reduction in recall (11%) in passing

from configuration A to B. As pointed out above, categorical lexical constraints cause competing scores to tie in TPs with postverbal nominal constituents, when the lexical entry of the verb is specified for both a transitive and an intransitive frame. Actually, a real improvement is observed when disambiguation is driven by probabilistic lexical constraints: without threshold, precision is 72.90% and recall 98.70%, with an overall f-score which is 4% higher than f-score of configuration B.

Results reported so far also include the noise of parsing errors. By focussing on "gold" TPs only, precision and f-score figures are significantly higher, reaching 95% of precision in the configuration C with T≥2. However, the best f-score value is associated with configuration C with no threshold, where a good balance is achieved between precision and recall. It is interesting to note that also in this case the high value of precision in configuration B (outnumbering the first two values of configuration C) is associated with the lowest recall value.

Discussion

Careful analysis of disambiguation errors in the different configurations shows that they are almost always associated with the postverbal position. The same holds for ties in configuration B, all of which occur in TPs with the nominal constituent in postverbal position.

We report here the results of the analysis of the 27 wrongly disambiguated TPs in configuration C with T≥2 together with an indication of the types of constraints – if any - that could be resorted to for the correct disambiguation.

1. about one third of disambiguation errors could be avoided if also distributional semantic constraints were considered. This is the case of TPs like *dichiarare-guerra* 'declare-war' where the strong preference of *dichiarare* (in its 'report' sense) for taking postverbal subjects should be counterbalanced by the semantic oddness of *guerra* as a subject of *dichiarare*;
2. another third of errors concern TPs where both interpretations are, for lack of further context information, equally plausible, e.g. *raggiungere-Germania* 'reach_Germany';
3. for some of the problematic TPs in 2), knowledge about the wider syntactic context where they occur is required: this is the case, for instance, of TPs occurring in wh-clauses, where other factors appear to combine with the lexical ordering constraints successfully used for TPs in main clauses;
4. in a few cases, presence/absence of a (definite/indefinite) article can be used as a disambiguating constraint (see TPs like *occupare-posto* or *dichiarare-guerra* where the nominal constituent lacks an overt article);
5. finally, there are ambiguities which would not arise if domain-specific lexicons were available: e.g. in the TP *interessare-risultato* 'interest-result' the noun is used as a direct object; however, in our training corpus this transitive use of *interessare* seems to be restricted to the financial domain only, while our testbed was sampled from a general reference corpus. In fact, different verb senses tend to have different subcategorization probability (Roland and Jurafsky, 2003). Hence, careful estimation of the latter should take into account domain specific senses with a bias towards a particular type of frame.

The examples discussed above call for further refinements and extensions of the constraint module of IDEAL+, ranging from semantic distributional constraints (currently being tested) to domain-specific lexical constraints and further granular constraints on article use or constituent ordering in specific syntactic constructions (e.g. wh-clauses).

Conclusions

We presented a battery of experiments where various configurations of constraints have been applied to a specific parsing subtask, i.e. subject/object disambiguation. Results confirm the key-role of lexical information for grammatical dependency identification, provided that this information is used probabilistically. For instance, the fact that a large number of Italian verbs (almost ¼ of the whole lexicon) occur both with a transitive and intransitive frame has a negative impact on the disambiguation process, unless we have a clear estimation of the verb preferences for either type of reading. As a further aspect of interest, probabilities of lexical argument position can be combined usefully with more standard syntactic slot distributions. This may also suggest that lexical preferences on argument order should play a prominent role in human lexical knowledge (at least in languages, like Italian, with a rather free constituent-order). Although the present experiments focus on a specific task, obtained results encourage us to extend the approach to other parsing tasks - such as PP-attachment -, as well as to augment the constraints module by combining syntactic and semantic information

References

Bartolini, R., Lenci, A., Montemagni, S., Pirrelli, V. (2002), "Grammar and Lexicon in the Robust Parsing of Italian. Towards a Non-Naïve Interplay", in *Proceedings of the International COLING-2002 Workshop "Grammar Engineering and Evaluation"*, Taiwan.

Basili, R., Zanzotto, F. M. (2002), "Parsing Engineering and Empirical Robustness", *Journal of Natural Language Engineering*, VIII.

Manning, C. (2003), "Probabilistic syntax", in Bod, R. *et al.* (eds.), *Probabilistic Linguistics*, MIT Press, pp. 289-341.

Montemagni, S. (1995), *Subject and Object in Italian Sentence Processing*, PhD Thesis, University of Manchester Institute of Science and Technology (UK).

Montemagni, S. *et al.* (2003), "Building the Italian Syntactic-Semantic Treebank", in Anne Abeillé (ed.), *Building and using Parsed Corpora*, Language and Speech series, Kluwer, Dordrecht, pp. 189-210.

Jurafsky, D., Martin J.H. (2000), *Speech and Language Processing*, Prentice-Hall.

Roland, D., Jurafsky, D. (2002), "Verb sense and verb subcategorization probabilities", in Stevenson, S. and Merlo P. (eds.), *The Lexical Basis of Sentence Processing: Formal, Computational, and Experimental Issues*. Amsterdam, John Benjamins, pp. 325-346.