How NLP techniques can improve speech understanding: ROMUS – a Robust Chunk based Message Understanding System Using Link Grammars

Jérôme Goulian, Jean-Yves Antoine, Franck Poirier

VALORIA Laboratory, University of South-Brittany
Yves Coppens Research Center, Tohannic
F-56000 VANNES, France
jerome.goulian@univ-ubs.fr

Abstract

This paper discusses the issue of how a speech understanding system can be made robust against spontaneous speech phenomena (hesitations and repairs) as well as achieving a detailed analysis of spoken French. The *Romus* system is presented. It implements speech understanding in a two-stage process. The first stage achieves a finite-state shallow parsing that consists in segmenting the recognized sentence into basic units (spoken-adapted *chunks*). The second one, a Link Grammar parser, looks for inter-chunks dependencies in order to build a rich representation of the semantic structure of the utterance. These dependencies are mainly investigated at a pragmatic level through the consideration of a task concept hierarchy. Discussion about the approach adopted, its benefits and limitations, is based on the results of the system's assessment carried out under different linguistic phenomena during an evaluation campaign held by the French CNRS.

1. Introduction

Parsing spontaneous speech is a difficult problem. In the area of speech understanding in task oriented Human-Machine Communication, the input is perturbated by two influences: the false starts, self-repairs and ungrammatical constructions that are produced by the speakers and the recognition errors. Most work so far has therefore primarily focused on dealing with narrow and well-defined domains. Indeed, when the domain is restricted, sufficient coverage can be achieved using semantically guided approaches that allow skipping of unparsable words or segments. The linguistic analysis is then restricted to a mere extraction of concepts and case markers [1]. For instance, the understanding component of speech information retrieval systems considers only the words needed to generate a database query. Although these approaches have led to significant results, two points are to be noted. First, despite their robustness on many of spontaneous phenomena, they mostly require some *ad hoc* mechanisms to deal with self-repairs or restarts [2]. Corrective methods based on robust pattern matching [3] or stochastic language models including lexical or acoustic knowledge [4] have been investigated as a preliminary stage process. Although they give interesting results they fail to detect and properly handle all type of corrections and need incorporate higher level syntactic and semantic processing [4]. Secondly, the portability to other application domains of such approaches remains an open issue: one should reasonably assume that less restricted tasks require a more detailed linguistic analysis [5]. Whatever the answer to the latter question may be, this paper investigates whether linguistically sophistication in the analysis can improve speech understanding while preserving robustness. We argue it is possible to obtain more detailed representations thanks to a combination of a syntactically- followed by a semantically-driven robust parsing with a reasonable amount of effort. We will show we can capture therefore a majority of information that can be employed for disambiguation, detection and process of repairs, and that can be useful for context interpretation.

It was demonstrated that parsing natural language can be handled by incremental approaches based for instance on the principle of "chunking" the input into small and easily manageable units [6]. *ROMUS*, the speech understanding system we present in this paper, applies and sensibly adapts these ideas to spontaneous spoken dialogue. *ROMUS* achieves speech understanding in a two stage process. The first stage segments the utterance in grammatical units. These units are then attached together using a Link Grammar parser. The inter-chunks linkages are mainly investigated at a pragmatic level through the consideration of the task concept hierarchy. However, it takes into account some of the morphological cues that are captured during the first stage.

The task of the *Romus* system concerns tourist information (hotel reservations, asking for directions in town, etc.). This domain seems more complex than traditional restricted domains (ATIS for example). For instance, anaphora resolution should need a finer linguistic analysis to be correctly handled; multi-queries on different objects are allowed[1].

The organization of the paper is as follows: in section 2 we introduce the concept of chunk parsing, how we interpret it and use it in our system. Section 3 deals with the semantic Link-Grammar parser that it is used. We focus in particular on the process of the unexpected (hesitations, self-repairs). In section 4, results from the evaluation of the system within the GDR I3 program of the CNRS (French Research Agency) are presented and discussed.

2. The Chunker

Recent methods of NLP shallow parsing have shown their ability to process unrestricted text in a robust way [7]. Many of these lightweight techniques use finite-state parsing that offers the advantages of computational efficiency and the ability to easily integrate a number of level of processing

[1] One can asks for instance about name, address and prices of cinemas and museum near a given place.

[8]. In particular, our system adapts chunk parsing to spoken Human-Machine Communication using finite-state transducers from morpho-syntactic tagging to the segmentation itself. This frist stage, which is only based on syntactic considerations, produces generic segments and can be therefore considered as a first step towards portability issues of SLU[2].

Tagging The first step consists of looking up each word of the input in the lexicon to assign it syntactic labels. The lexicon is achieved by a deterministic finite-state automaton with both low access time and small storage space. This lexicon encodes 45 000 words. The set of distinct syntactic part-of-speech tags, that is used to segment the utterance, is composed of 34 distinct tags. Some of the morphological informations encoded in the lexicon are however retained. For instance the distinction between definite and indefinite articles is stored at this stage (as this information is crucial for anaphora resolution) but is not relevant for the segmentation purpose.

To deal with lexical ambiguity, a few contextual rules are used and encoded in finite-state transducers. Most of them are the ones used for written French lexical tagging as it seems to be possible to directly transpose them to spoken French [10][3]. As in robust text parsing, a low decision rate has been prefered[4]: Residual lexical ambiguity is handled with parsing with each possible sequence that remains.

Chunking The segmentation is based on these tags. In the context of spontaneous spoken dialog, three different groups of *chunks* (minimal and non-recursive constituents [6]) have been considered:

- regular syntactical groups: verbal chunks, nominal chunks, adjectival chunks, adverbial chunks and prepositional chunks. These groups consist of a single content word (the lexical head) surrounded by a constellation of function words, matching a fixed template. Verbal chunks never incorporate their arguments but include only the structural complexity (modal form for instance); nominal chunks never incorporate adjectives, etc. that follow the noun. Coordinations, relative pronouns, etc. are considered as "one-word" specific chunks.
- groups that correspond to domain-independent language expressions such as dates, times, prices, etc.
- finally, "speech" chunks include the markers of repairs, repetitions and other spontaneous constructions.

Each group above can be described by regular expressions that are compiled into non-deterministic finite-state transducers which are then determinised [11]. Each transducer is used as a finite state marker around the groups it defines in a cascaded way. Segmentation ambiguity is handled as Abney's parser [12] thanks to a left-to-right longest match parsing strategy. Figure 1 shows the sequence of chunks extracted on the correct tagged sequence of a french utterance. Ambiguity of part-of-speech tagging is reduced by half by the segmentation process[5]. It is to be noted that the syntactic chunk-parsing provides a preliminary treatment of some corrections: words that are not included within a chunk at this stage are removed. One of the regularities of spoken language is indeed that repetitions often restart at the beginning of the current syntagm the user wants to correct [13, 14]. This regularity has motivated our choice to consider prepositional chunks at this level.

3. The Link-Grammar parser

Each syntactical chunk can be viewed as a local dependency tree whose root is the lexical head of the chunk as shown figure 2. The second stage of the understanding process

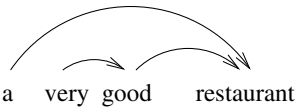

Figure 2: Example of the local dependency tree associated with the nominal chunk *a very good restaurant*.

achieves a lexicalized analysis that aims at characterizing dependency relations between chunks. Theses relations will correspond to the predicate-argument relations of the final semantic structure. Local dependencies that have been extracted from each chunk are not taken into account by the linkage process. They are stored and will complete the final semantic representation. By reducing the sentence to chunks, there are fewer units whose associations must be considered [12]. Each chunk is mapped onto a three element structure: the chunk category (C), the lexical head (H) and some morphological cues (M) that are relevant for the linkage process (typically presence of definite or indefinite articles, possessive adjective etc.) as shown in table 1. We consider both the preposition and the lexical head of the dependant group as lexical heads of prepositional chunks[6]. Global attachment between chunks is based only on the $< C, H, M >$ information.

The dependencies are mainly investigated at a pragmatic level through the consideration of the task's concept hierarchy[7]. We develop this parsing process in the framework of Link Grammar [17]. Link Grammar is a grammar formalism which is based on the words in the lexicon and their linking requirements. Links are labeled connectors which can attach to matching links to the right or left of the words. A sentence of the language defined by the grammar is a sequence of "correctly linked" words. For the *Romus* Link-Grammar parser, dictionary's entries are not words but each combination of C, H and M discussed above.

$$(E_1) < C_1, H_1, P_1 >: (R^- \; and \; \{T^+\});$$

[2]Note that this first stage is also used for the analysis of concepts which recur in many domains such as expressions of date or time for instance. The preliminary analysis of these concepts is also important for portability issues [9].

[3]Specific spoken French rules have been however designed to deal with words whose common use in spoken french language sensibly differ from the one in written language (the word *quoi* (*what*) for instance that is mostly used as an interjection in spoken french but still can be a relative or interrogative pronoun).

[4]The aim is The decision rate of the *Romus*' tagger, carried out on a set of 1200 utterances described in section 4, is 80.4%. Precision rate, carried out on a similar set of 1200 utterances, is 97.5%

[5]Sequences that provide the best local coverage are retained. For instance, for the sequence *la gare* (*the station*) where *gare* can be a noun or a verb, the nominal chunk is prefered as it includes the article.

[6]Since the linkage process is domain specific, the preposition alone may be sufficient. This is the case for *near the station* as any different semantic role of the preposition *near* in the application would have been allready detected: *near Christmas* for instance is a Prep. Date Chunk

[7]The task concept hierarchy has been derived from the analysis of the PARISCORP corpus (1400 utterances) [15]. *Romus* incorporates currently 36 standard queries (time, price, reservation, etc.) and 151 concepts (objects – e.g. "hotel" – and properties – e.g. "with shower" –). This can be compared with the ARISE domain that requires 8 queries and 64 concepts [16].

[je\$_{pr}$]GN [cherche\$_{vb-conj}$]GV un\$_{indefinite-det}$ [un\$_{ind-det}$ restaurant\$_{n}$]GN[euh\$_{hes}$]HES
[un\$_{ind-det}$ restaurant\$_{n}$]GN [chinois\$_{adj}$]GAdj près-de\$_{prep}$[près-de\$_{prep}$ (la\$_{def-det}$ gare\$_{n}$)GN]GP

Figure 1: Extracted chunks on the French utterance "*I'd like to find a a restaurant hum a chinese restaurant near near the station*". Syntactic tag is noted after backslash. Chunks are represented between brackets.

Chunk	C Category	H Lexical Head	M Morphological Cues	D
want to know	Modal-Verbal	to know		*
a restaurant	Nominal	restaurant	indef.	*
near the station	Prep.	near	def.	*
for children	Prep.	(for, child)		*
Christmas	Date	$\emptyset$	$\emptyset$	
before 5 o'clock	Prep. Time	before	$\emptyset$	*
and	Coordination	and	$\emptyset$	
sorry	Correction marker	sorry	$\emptyset$	

Table 1: Examples of $< C, H, M >$ information assigned to each segment and involved in the linkage process. A * in the D column indicates presence of local dependencies.

$$(E_2) < C_2, H_2, P_2 >: R^- \; or \; R^+;$$

Theses two simplified entries mean for instance that E_1 can satisfy an R relation to another element at its left in the utterance (operator$^-$) and optionally a T (curly brackets) relation with another element at its right (operator$^+$). The succession of E_2 E_1 in an utterance is therefore a "valid" utterance expressed by the R relation that links the two elements together.

One of the main challenge of this step is to deal with the flexibility needed both by the dysfluencies and by the weak word-orderfreedom of spoken French. Word-order linguistic phenomena can be handled in this framework thanks to the important panel of operators that can be used in the link requirements expression. Direct automatic modelling of such phenomena is provided. Furthermore partial analysis is authorized: the two conditions that have to be satisfied are the linking requirements of each element of the sequence and the fact that links can't cross (non-projective utterances are therefore prohibited).

Two principal levels of relations have to be distinguished: domain dependent relations[8] and the ones that handle syntactical constructions such as logical coordinations. Marked repairs and repetitions are treated similarly as logical coordinations. Link requirements are governed by the logical coordination or the marker of correction thanks to sets of semantic compatible relations. Dictionary's entries for the markers are as follows:

$$< Coo, and, \emptyset >: (COO^-_{X_1} \; and \; COO^+_{X_2}) \; and \; X^-;$$

where $X1, X2 \subset X$ denote *AND-compatible* links. Figure 3 shows linkage computed according to this rule. In this example, the *Room-Category* and *Room-Quality* relations are *AND-compatible*. Two *Room-Quality* are not *AND-compatible* (a room has either a shower or a bath but never both). On the contrary, they are *NO-compatible* to express a correction with the *no* marker. *AND-compatible*, *NO-compatible*, *OR-compatible* relations are defined from the semantic description of the domain provided by the analysis of the task universe.

[8]Theses relations express *argument* relations (an hotel for instance can be the object of a rate query) and *property* relations such as *Room-Category* that can link a *room* (of a hotel) to the segment *double* for instance.

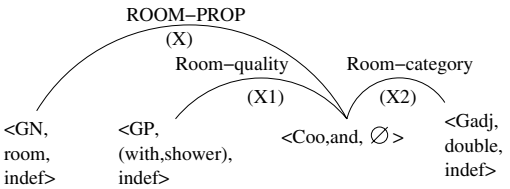

Figure 3: Example of coordination computed links.

Our dictionary contains currently about 1000 entries. Reduction of the entries considered is achieved by mapping lexical heads of segments to the word that represents the specific categorization of the segment within the domain. An iterative and semi-automatic process is used to generate the dictionary from a semantic grammar that describes the domain application.

The parsing algorithm with $O(n^3)$ runtime[9] is adapted from the one given by [17]. Each candidate linkage receives a cost. Cost 0 correspond to complete linkages (all segments link with each other). The cost is then proportional to the number of segments ignored. Linkages with islands[10] are therefore preferred. An additional cost lets prefer linkages where the sum of link-distance between segments is the lowest.

The final semantic representation is then built in exploring the non-oriented graph that is obtained. At this stage non-marked repetitions and repairs, as well as multi-queries, are treated by the domain-dependent compatibility rules mentioned above. If no decision can be made, the ambiguity is stored in the representation. We can consider for instance the example of figure 3 without the *and* marker. If no more markers appear in the rest of the utterance, we do not decide whether it is a correction or an enumeration as the two relations are both *AND-compatible* and *NO-compatible*. On the contrary, in the following example "*room with shower with bath*", a correction is detected as the two relations are not *AND-compatible*. The second segment is then stored in the semantic representation. Finally, the semantic

[9]n is the number of chunks in the utterance; it is rarely higher than 12.

[10]An island correspond to at least one linkage between two segments that are not linked with the rest of the utterance.

representation includes any of the information (morphological, syntactical, ambiguity) that were captured during process. Some of them (as demonstrative adjectives for instance) can then be used for contextual interpretation.

4. Evaluation

Romus has been assessed during the evaluation campaign founded by the GDR I3 (Intelligence-Interaction-Information) program of the CNRS french Research Agency[11]. This objective evaluation aims at providing a detailed diagnosis of the behaviour of the involved understanding systems[12] on specific phenomena. Each system has been evaluated on a specific set of 1200 tests sentences, each incorporates a specific problem. Three principal different kinds of difficulties have been evaluated: spontaneous speech (incidental clauses, false starts, repetitions and corrections), linguistic phenomena that involve modification of standard word order and complex utterances (multi-queries or multi-objects queries). In this section, *Romus*' results are presented; details and global results of the campaign can be found in [18]. Three scores have been used to evaluate global performance of the system: G (Good) is the number of utterances that led to a complete and correct semantic representation, K (oK) is the number of utterances that led to an incomplete semantic representation but preserve the global meaning (constraint or property deletion for instance), and B (Bad) is the number of utterances that led to an incorrect semantic representation. Table 2 presents *Romus*' error rate on each class of phenomena[13].

Tested phenomena	*Nb tests*	K	B	% error
Spontaneous Speech	390	22	7	7.4
Segment free order	293	8	7	5.1
Complex utterances	196	4	6	5.1
Combined phenomena	96	18	20	39.5

Table 2: *Romus*' Error Rate on each tested phenomena.

The results are promising in dealing with spontaneous dysfluencies, complex utterances and modification of segment order. The error rate observed on utterances that incorporate combined phenomena is quite promising. Tests include indeed a large set of combined spontaneous and word-order phenomena and therefore correspond rather to human-human dialogues than to human-machine dialogues. Errors mainly occur on segment free order combined with complex structures (more than two coordinations and corrections). Our system is currently evaluated more precisely in the French MEDIA evaluation campaign held by the French research Ministry (TECHNOLANGUE projects).

Global results of our system is similar to the others. Potentials of this approach reside however in the genericity of the chunk parsing process and the high level of accuracy of the semantic representation that can be produced. Moreover, thanks to the combination of syntax and semantics, this strategy do not require any prelimary treatment of repetitions and corrections that appear to led to spurious correction.

[11] This campaign had concerned five french laboratories : VALORIA, CLIPS, IRIT, LIMSI and LORIA.

[12] This campaign had concerned four laboratories.

[13] % of error is obtained by $\frac{B+K}{G+K+M}$

5. References

[1] W. Minker, A. Waibel, and J. Mariani, *Stochastically based semantic analysis*. Amsterdam: Kluwer, 1999.

[2] S. Issar and W. Ward, "CMU's Robust Spoken Language Understanding System," in *Eurospeech'93*, 1993, pp. 2147–2150.

[3] J. Bear, J. Dowding, and E. Shriberg, "Integrating multiple knowledge sources for detection and correction of repairs in Human-Computer dialogue," in *ACL'92*, Newark, Danemark, 1992, pp. 56–63.

[4] P. Heeman and J. Allen, "Improving robustness by modeling spontaneous speech events," in *Robustness in language and speech technology*. Dordrecht, Pays-Bas: Kluwer Academic Publishers, 2001, pp. 123–152.

[5] G. van Noord, G. Bouma, R. Koeling, and M. Nederhof, "Robust Grammatical Analysis for Spoken Dialogue Systems," *Natural Language Engineering*, vol. 5, no. 1, pp. 45–93, 1999.

[6] S. Abney, "Parsing by chunks," in *Principle Based Parsing*. Kluwer Academic, 1991.

[7] J.-P. Chanod, "Robust Parsing and Beyond," in *Robustness in Language and Speech Technology*. Dordrecht, Pays-Bas: Kluwer Academic Publishers, 2001, pp. 187–204.

[8] E. Roche and Y. Schabes, Eds., *Finite state Language Processing*. MIT Press, 1997.

[9] S. Seneff, "The use of linguistic hierarchies in speech understanding," in *ICSLP'98*, Sydney, 1998, pp. 3321–3333.

[10] A. Valli and J. Véronis, "Etiquetage grammatical des corpus de parole : problèmes et perspectives," *RFLA*, vol. IV, no. 2, pp. 113–133, décembre 1999.

[11] E. Roche and Y. Schabes, "Deterministic Part-of-Speech Tagging with Finite State Transducers," in *Finite state Language Processing*. MIT Press, 1997, pp. 205–239.

[12] S. Abney, "Partial parsing via finite-state cascades," in *Workshop on Robust Parsing, ESSLLI'96*, 1996, pp. 8–15.

[13] C. Blanche-Benveniste, *Le français parlé ; études grammaticales*. Paris: CNRS Editions, 1990.

[14] W. Levelt, "Monitoring and self-repair in speech," *Cognition*, vol. 14, pp. 41–104, 1983.

[15] S. Rosset, L. Lamel, S. Bennacef, L. Devillers, and J.-L. Gauvain, "Corpus oral de renseignement touristique," in *Ressources et évaluation en ingénierie des langues*. De Boeck Université, Duculot, 2000, pp. 483–489.

[16] H. Maynard and F. Lefèvre, "Apprentissage d'un module stochastique de compréhension de la parole," in *JEP-TALN'02*, Nancy, France, 2002.

[17] D. Sleator and D. Temperley, "Parsing English with a Link Grammar," CMU-CS-91-196, Tech. Rep., 1991.

[18] J.-Y. Antoine, C. Bousquet-Vernhettes, J. Goulian, M. Kurdi, S. Rosset, N. Vigouroux, and J. Villaneau, "Predictive and objective evaluation of speech understanding." in *LREC 2002*, 2002.