Shallow Parsing and Information Extraction

Diego Uribe

Departamento de Ingenierías
Universidad Iberoamericana-Laguna
diego.uribe@lag.uia.mx

Abstract. From the set of modules that a full IE system comprises, the component to deal with the definition of text patterns for the identification of specific relational information has captivated the attention of researchers. Much research has been carried out in the context of subject/object relationships for the identification of the protagonists involved in target text events. Relatively little has been reported on learning rules for text extraction based on chunks as the main syntactic unit. In order to analyze the plausibility of an intermediate granularity such as chunks, our model implements a rule representation which has proved to be both effective and flexible. Regardless of the number of parameters (slots) the event represents, the rule representation is able to denote the local context which surrounds the target fillers. Experiments conducted in two different domains give evidence of the adaptability of the model and linguistic analysis of the corpus was carried out in order to illustrate how IE is a domain-specific task. We demonstrate how not only tuning the semantic lexicon to a new scenario plays a key role in the adaptation process but also how the syntactic analysis gets into trouble when the partial parser is confronted by a specialized corpus.

1 Introduction

Information Extraction (IE) is the activity concerned with the extraction of specific information through skimming natural language text. Actions or events and their protagonists commonly known as entities describe this target information. The basic architecture of this activity is made up of three modules: preprocessing, syntactic analysis and domain analysis [1]. While the first two modules are domain independent, that is, the preprocessing and syntactic analysis of the text can be implemented regardless of the concerned domain, the last module, domain analysis, is a module that draws on the knowledge which surrounds a particular area.

An important aspect in the learning process for the automatic detection of particular events from plain text documents is the rule representation. A rule representation based on subject-verb-object relations is a common scenario to facilitate the identification of the main protagonists in the target event. However, a more fine-grained way of representation such as noun groups and verb groups alters the search space, so that a suitable similarity metric and generalization procedure are demanded by the learning mechanism. In this paper, we approach

© A. Gelbukh, R. Monroy. (Eds.)
Advances in Artificial Intelligence Theory
Research on Computing Science 16, 2005, pp. 45-54

this problem through string similarity. Specifically, we turn this problem into the *longest common subsequence* problem in order to find the longest common subsequence among the chunks which constitute the training instances.

The analysis of two entirely diverse scenarios allows us to show two more interesting implications of a rule representation based on an intermediate granularity. First, the grammatical variation in the narration of the target event plays a key role in the identification of the event's parameters. While a common practice to cope with this problem is to make use of normalization [2], we can see how some scenarios do not require such approach. Secondly, the difficulties that the partial parsing exhibited with a specialized domain give evidence to suggest that the module of syntactic analysis is also domain dependent.

In addition to the syntactic elements which the rule representation is based on, the number of fillers that the learned rule is able to extract is another important aspect of the rule representation. According to the studied scenarios, a flexible structure which allows the identification of the context that surrounds the relevant phrases is mandatory. For this purpose, an extension of the representation implemented by RAPIER [4] seems worth of experimentation.

Finally we describe the construction of the domain-specific glossary known as semantic lexicon. In fact, the description of the specific sub-language for our particular domain relies on both the meaning of the words and the meaning associated with our syntactic constituents: noun and verb groups. In fact, the set of semantic word classes which make up the lexicon may be classified according to the content word represented by noun groups and verb groups: while noun groups are used for the identification of either the major objects or facts in the domain, the most important actions or relationships are spotted by the analysis of the verb groups.

In this way, this paper will argue that a through linguistic analysis of the corpus is needed to get around the peculiarities that a specialized domain entails. In other words, rather than considering the domain analysis module as the only domain dependent module, we show how the three modules which make up the basic architecture of IE demand a rigorous linguistic scrutiny.

In the next section we give a brief description of the studied scenarios. In section 3 we review the problem of a similarity metric for a rule representation based on chunks. In section 4 we show how the grammatical variation and the scattering of events in text are relevant properties of the scenario. We also show in this section, examples of how the partial parser struggled with a specialized scenario. In section 5 we describe how the definition of the lexicon is beyond a straightforward analysis of the content words. Finally, we conclude with a discussion in section 6.

2 Scenarios

For our experimentation, we have chosen two scenarios. One is a "classic" data set: Management Succession [3], and the other one is a specialized corpus: the football domain. Since the former corpus is a very well known scenario, we describe in this section the football domain. The corpus consists of a set of articles, commonly known as "clockwatch", which describe the main events of a football match: goals, substitutions, booked players, missed chances, etc. This collection of articles has been drawn from the BBC Sport website[1]. Our corpus contains 101 texts and, the average size is roughly 1000 words per match-description (109,370 words in total results in an average of 1082 words per text).

At first glance, our corpus seems quite simple but the linguistic complexity of the events is beyond shooting a ball. The linguistic complexity of the football texts is as worthy of linguistic analysis as those that were used in the Message Understanding Conferences, namely: Terrorism in Latin America, Join Ventures activities and Management Succession events. According to the five information extraction tasks, as defined by Cardie [5], football texts provide enough material for linguistic research for each of these tasks. For example, Named Entity Recognition is an imperative task for the identification of the protagonists involved in the events, that is, players, managers, referees and places as stadiums. There are also goal descriptions in which Coreference Resolution would be very useful for the elucidation of anaphoric references (frequently pronouns) that permit to identify the corresponding protagonists (or antecedents) in the event. The identification of relations between the entities, known as Template Relation construction, is another essential task that would be also very convenient for the recognition of the players involved, either when a goal is scored or when an incident has occurred. So, as we can see within the vastness of events (along with their corresponding peculiarities) that surrounds football, this domain also represents a significant linguistic challenge.

Table 1 shows some sentences which represent the set of target events: goals, substitutions and booked players; and the corresponding filled templates for each of the sentences are shown in table 2. In these examples we can note how the protagonists for each event play a specific role and denote a relationship between them. For instance, the template for the *goal* event illustrates a relationship between two players, the passer (PasserPlayer) and the striker (StrikerPlayer), when a goal is scored. In an analogous way, the template for the *substitution* event illustrates a substitution relationship between two players, the player who comes in (InPlayer) and the player who goes off (OffPlayer). The next event, *booking*, illustrates an offensive relationship between a player who is the aggressor (BookedPlayer or ExpelledPlayer) and another player who is the offended (InjuredPlayer) or the cause (Reason) for which the aggresor has been booked with either a yellow or red card.

3 Rule Representation

Given that, on the one hand, an event alludes to an action or a relationship between entities, and on the other hand, we are interested in extraction rules based on local context only, that is, multiple-sentence event descriptions are

[1] http://news.bbc.co.uk/sport1/hi/football/default.stm

Table 1. Four sentences which illustrate the target events

Goal
20:58 Juan Sebastian Veron plays a fabulous through-ball to Ole Gunnar Solskjaer, who takes a touch before firing under Deportivo's keeper from an acute angle.
Substitution
78 mins: The outstanding Gerrard is substituted for Bayern Munich's young midfielder Owen Hargreaves.
Booking
9 mins: England midfielder Scholes is booked for a reckless late tackle on Charisteas.

Table 2. Filled templates

Goal	`Minute`	`PasserPlayer`	`StrikerPlayer`
	20:58	*Juan Sebastian Veron*	*Ole Gunnar Solskjaer*
Substitution	`Minute`	`InPlayer`	`OffPlayer`
	78	*Gerrard*	*Owen Hargreaves*
Booking	`Minute`	`BookedPlayer`	`InjuredPlayer/Reason`
	9	*Scholes*	*Charisteas*

beyond the scope of this work, we must look for a frame which allows us to stand for the local context that encloses an event.

Among the multiple concept representations which have already been observed in the past (most of them based on frames in terms of classic syntactic constituents), we have devised a rule representation which synthesizes the features of two previous systems. LIEP's representation [6] describes the syntactic context of the path between the target entities, that is, the slot fillers represented as noun groups, as well as the semantic classes of the heads of the target nouns groups. Second, RAPIER's representation [4] denotes both syntactic and semantic information around a target noun only. In this way, a plausible rule representation is the integration of these previous abstractions. Thus, in our system, the linguistic constituents have been framed taking into account both schemes: the path between the target entities and the context that surrounds the entities to be extracted.

In fact, the local context that surrounds an event is represented by a structure which contains four components (or fields): *pre-filler*, *filler*, *link* and *post-filler*. *Filler* represents one of the target entities, that is, a slot in the template. *Pre-filler* corresponds to the text that precedes the first filler of the event, whereas *post-filler* denotes the text after the last filler. Finally, *link* is concerned with the path between the event's fillers. Thus, an action or event of n fillers is represented by $2n+1$ components or fields. The next example: "1617 Arsenal continue to punish Fulham as Thierry Henry scores his second goal from another pass from Sylvain

Wiltord.", is an instance of a goal event, and the corresponding representation is shown in table 3.

Table 3. Syntactic and semantic elements for an instance

Element	Chunk list
Pre-Filler	[1617_CD]] [[Arsenal_NNP]Team] ((continue_VBP)) ((to_TO punish_VB)) [[Fulham_NNP]Team] as_IN
Filler1	[[Thierry_NNP Henry_NNP]Player]
Link	((scores_VBZ)Score) [[his_PRP second_NN goal_NN]Goal] from_IN [[another_DT pass_NN]Pass] from_IN
Filler2	[[Sylvain_NNP Wiltord_NNP]Player]
Post-Filler	._.

A deeper analysis of the information in this table is worth our attention. First, we can see the syntactic and semantic constituents corresponding to the event represented by a specific training instance[2]. The plain text is submitted to a shallow syntactic processing known as chunking which produces a flat list of noun groups, verb groups and function words. These in turn will be processed by the entity recognizer and the semantic tagger respectively in order to provide an elementary interpretation of the syntactic constituents. Secondly, we note the particular components that make up the representation of a goal event. This particular structure contains two fillers, the *link* field between them and the *pre-filler* and *post-filler* fields respectively; therefore, the total number of components in the structure which represent this event is five.

3.1 Similarity Metric

How LEEP discerns the similarity between two positive training instances is a question that revolves around the underlying rule representation. More exactly, the granularity that underlies the meaning representation plays such a relevant role that a suitable criterion for similarity must be devised. For example, previous rule extraction learning systems for text such as AutoSlog [8] and Crystal [9] make use of unification as the computational technique for their similarity purposes according to their rule representation defined in terms of conventional syntactic elements as subject, object, preposition phrase, etc.

Given that the number of conventional syntactic categories is small and well defined, the use of unification as computational technique for similarity purposes seems appropriated for rule extraction learning systems for text whose rule representation is based on these conventional features. However, a finer grained meaning representation, as a result of the partition of a sentence into smaller

[2] Round brackets represent verb gropus whereas square brackets stand for noun groups. Word classes are shown between brackets.

linguistic units, that is, grounded in terms of chunks, gives rise to an undefined and less precise definition of the number of syntactic features to be analyzed. In fact, when the number of linguistic units to be considered in the similarity processing has been not only increased but also undefined, that is, the number of features to be considered is variable, the evaluation of the distance between two training instances is more intricate.

The challenge then is to find out a similarity metric that allows us to identify regularities among the linguistic units in which our rule representation is based on: noun groups, verb groups and function words only. Since the meaning of a sentence is based not only on the words that make it up, but also on the ordering and relations among the words [7], we try a similarity metric which takes into account the ordering among the linguistic features: *longest common subsequence.*

To accurately distinguish the protagonists involved in the event, previous systems such as AutoSlog [8] and Crystal [9] regard as a similar instance that one whose slots to be extracted must be contained in the same syntactic constituents as the seed's slots. For example, if the seed represents a substitution event such as *Owen is replaced by Heskey* in which *Heskey* represents the player who comes in and *Owen* the player who goes off, an event such as *Sven-Goran Eriksson replaces Gerrard with McManaman* could not be a candidate instance worth having similarity analysis, since the fillers to be extracted do not belong to the same syntactic constituents (while the subject in the former sentence represents a target player, the subject in the latter sentence represents a different concept: the manager).

But for us, instead of looking for similarity among specific syntactic constituents, we care about the order of the slots to be extracted as an effect of the rule representation implemented. By making use of the same example shown in the previous paragraph, the event *Sven-Goran Eriksson replaces Gerrard with McManaman* could be considered a candidate instance to similarity evaluation, as the order of the target concepts is the same than the seed. However, an event such as *Kanu replaces Thierry Henry* could not be a similar instance since even though it describes a substitution event, it does not exhibit the same order in the fillers (*Thierry Henry* is the player who goes off, whereas *Kanu* is the player who comes in.). In this way, in order to distinguish between the protagonists involved in an event, we are concerned with a crucial linguistic element in its description: the voice. To cope with this grammatical variation, inflectional information about specific verbs is included in the semantic lexicon. In fact, sets of verbs which make up a particular semantic class are represented by the inflected form of the class. For example, verbs such as *replace* and *substitute* are represented as “Replace” or “Replaced” according to the voice of the event.

4 Syntactic Analysis

4.1 Grammatical Variation

The boundless expressive power of the language is present in the Management Succession domain. The most straightforward way to notice this inherent characteristic of the language is the scattering of events in text. In fact, the spread of the target information is a common scenario in this domain (multiple sentences describe the event) as well as the number of combinations which describe the relative order of the fillers in the expression of the target concept (single sentence or local context). For example, in the particular case of instances which contain information about Organization, PersonIn and Position, there are multiple combinations which must be considered for the definition of the corresponding patterns.

Even tough a wider variability in the narration of the target event in the Management Succession domain was experimented, the football domain also gave evidence of this phenomenon. How did our system cope with this grammatical variation? What were the implications of the use of an intermediate linguistic granularity?

Grishman [2], in order to surpass the limitation of partial parsing, that is, the incapacity to capture paraphrastic relations between different syntactic structures, defines metarules which expand subject/object relationships into patterns for the active, passive, relative, etc. clauses. In our particular case, we do not make use of metarules for the expansion of subject/object relations since these syntactic constituents are unknown for us. Rather than to define a pattern for each syntactic structure, we learn rules according to the arrangement of the fillers on the training instances corresponding to each target event. In this way, instances whose syntactic structure is *"X succeeds Y"*, or *"X, who succeeded Y"* are put into the same training set, whereas *"Y was succeeded by X"* belongs to a different set. Indeed, this is the way in which our system, according to the specific arrangement of the fillers rather than to a particular syntactic structure, finds out the salient regularities of the target events.

4.2 Partial Parsing

An important difference between the domains examined by our system was the fact that the MUC-6 texts did not represent any problem for the chunker. There were no practically syntactic errors in the processing of this generic type of texts, as opposite to the football texts which can be considered as texts which use a specialized lexicon.

In fact, as an indication of the complexity linguistic in the processing of the football domain is some parsing errors in which the chunker has incurred, particularly in goal declarations compared to the precision exhibited in the parsing of substitutions and cautioned player utterances. For example, in instances such as *"Steven Gerrard equalizes for Liverpool"*, *equalize* is considered as noun; in *"Nigel Martyn made a superb save to deny Scholes"*, *save* is sometimes tagged as verb or preposition.

5 Domain Analysis

The description of the specific sub-language relies on both the meaning of the words and the meaning associated with grammatical structures. In this way, determining the semantic class of a relevant noun group or verb group revolves around scrutiny of their corresponding heads. Nevertheless, the simple analysis of the heads is not enough: modifiers such as adjectives and adverbs may alter the concept so these elements also play a crucial role in the semantic tagging.

For example, in the football domain, lexical terms such as *almost*, or *deflected*, alter substantially the meaning of a head considered in isolation. Table 4 shows how the adverb *almost* acts on the verb *score* in the verb group *almost scores* (so in this case the POS pattern is "Adverb Verb"), whereas in the noun group *deflected shot*, the noun *shot* is modified by the adjective *deflected*.

Table 4. Modifiers such as *almost* and *deflected* in a verb and noun group respectively

80 mins: United sub Michael Stewart almost scores a spectacular goal but is denied by a fine Ricardo save.
The Scots were inmediatly on the offensive and a deflected shot from Stilian Petrov flashed narrowly wide.

There are, however, other situations in which modifiers are not always in close proximity to the head of the noun or verb group as in the previous case. In these circumstances, we made use of the variance-based co-ocurrence discovery approach introduced by [10]. Under this approach, important combinations of words for our corpus such as *shoots-wide* exemplify the vital role played by the modifier *wide*. The next instance *2022 Fabio Rochemback shoots just wide from 20 yards for Barcelona...* is an example which illustrates this case.

We also identified handy collocations in this domain. Collocations such as *make way for* represent a specific and concrete meaning which can not be captured by the semantic generality represented by the verb *make*. The next instance *75 mins Van Nistelrooy makes way for Diego Forlan after suffering a slight knock which requires an ice pack* is an example of an event described in terms of this collocation that comprises more than one chunk.

Finally, dealing with lexical ambiguity is also part of semantic interpretation. For example, the most important word in our football-corpus, *goal*, is ambiguous between two senses: the event which represents to score, and an area on a playing field, usually marked by two posts with a net fixed behind them (see table 5). Even though we are disregarding another possible meaning for goal: aim or purpose, it would not be possible for our semantic tagger to label this ambiguous term with a specific semantic category. In other words, the phenomenon of *polysemy*: a single lexeme with more than one meaning, is also present in our football domain. How did we cope with this situation? We dealt with this ambiguity case by making use of word collocations through statistical analysis of our corpus [11], specifically the frequency-based approach. By using this straightforward statistical method, we could work out a regularity that allowed us to distinguish the first sense (to score) of this term: an ordinal number followed by the term goal. Put in another way, in this corpus is very common to come across expressions such as *third goal* (Most of these expressions mean a goal for a specific team, but however, there may be expressions such as: ... on Derby with his 16th goal of the season, which mean a goal for a specific player rather than for a team.).

Table 5. Illustration of lexical ambiguity of goal

Sense	Instance
Score	83 mins Sydney Govou scores a vital goal for Lyon in Group D.
Area	4 mins Aristizabal hits the post when one-on-one with Mexican keeper Perez Rojas and just seven yards out from goal.

Likewise, more cases of polysemy were manually discovered during the annotation process of the texts. Metaphoric expressions such as *drill* and *slot* to describe the action when the player scores a goal, as well as terms such as *fire* and *rifle* to describe the action when the player shoots the ball, prove the great diversity of expressions used in the description of the football events.

6 Discussion

In the football domain, the experimentation conducted by our system suggests a deepest research in the recognition of the wide variability of terms used in the description of the events (lexical variation), rather than the use of subject/object relations to generate syntactic variants of a pattern, as a more convenient alternative to improve the covering of the learned rules.

Two particular difficulties were detected for the correct identification of specialized terms. First, the use of metaphoric expressions is present no only to describe the action to score (for example, terms such as *goes-off* in *Darius Vassell collects the ball and hits a superb right-foot shot that goes in off the post*, but also to portray the relationship between the players. For example, the use of *read* to describe the connection is illustrated in: *Berbatov read Yildiray Basturk's angled pass and slipped the ball across Barthez to reduce the deficit.* Secondly, the parsing problems shown in section 4 give evidence of how syntactic analysis is a module that goes through limitations when the parser deals with a specialized domain. Training the parser is a paramount activity to cope with this sort of domains.

On the other hand, the story is rather different for the MUC-6 corpus. The main difficulty in this domain is the grammatical variation. For example, a close scrutiny of the texts was required since a considerable percent of the information in the filled templates was obtained beyond local context: event descriptions spread across multiple sentences [12]. Furthermore, in order to cope with the wide diversity in the expression of the target concept (the potential number of combination among the template's fillers), the use of metarules is rather than essential in this domain. In this way, to some extent, an improvement in coverage is plausibly expected.

In this way, our analysis suggests that a deep linguistic analysis is essential to adapt our modules to extraction tasks. Two domains give evidence of how the complexity of a scenario relies on the peculiarities of its expressions [13].

References

1. Appelt, D., Israel, D.: Introduction to Information Extraction Technology. A Tutorial Prepared for IJCAI-99, (1999)
2. Grishman R.: The NYU System for MUC-6 or Where's the Syntax? In Proceedings of the Sixth Message Understanding Conference Morgan Kaufmann Publishers, 167-175, 1995.
3. MUC-6: Proceedings of the Sixth Message Understanding Conference San Francisco, CA, Morgan Kaufmann Publishers, 1995.
4. Califf, M.E., Mooney, R.: Relational Learning Techniques for Natural Language Information Extraction. The University of Texas at Austin, (1998)
5. Cardie, C.: Empirical Methods in Information Extraction. AI Magazine 39(1):65-79, 1997.
6. Huffman, S.: Learning Information Extraction Patterns from Examples. Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Learning Processing. Springer, 246-260, 1996.
7. Jurafsky, D., Martin, J.H.: Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice-Hall, Inc., 2000.
8. Riloff, E.: Information Extraction as a Basis for Portable Text Classification Systems. Ph.D. thesis. University of Massachusetts, Amherst, 1994.
9. Soderland S.: CRYSTAL: Learning Domain-specific Text Analysis Rules. University of Massachusetts, Amherst, (1997)
10. Smadja, Fraank: Retrieving collocations from text: Xtract. Computational Linguistics, Vol. 19, 143-177, 1993.
11. Allen, J.: Natural Language Understanding. The Benjamin/Cummings Publishing Company, Inc., Second Edition, 1995.
12. Stevenson, M.: Information Extraction from Single and Multiple Sentences. In Proceedings of the Twentieth International Conference on Computational Linguistics (COLING-2004), Geneva, Switzerland.
13. Huttunen, S., Yangarber, R., Grishman, R.: Complexity of Event Structure in IE Scenarios. In Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002).