Semantic Case Role Detection for Information Extraction

Rik DE BUSSER and Roxana ANGHELUTA and Marie-Francine MOENS
Interdisciplinary Centre for Law and IT
Katholieke Universiteit Leuven
Tiensestraat 41
B-3000 Leuven, Belgium
rik.debusser, roxana.angheluta, marie-france.moens@law.kuleuven.ac.be

Abstract

If information extraction wants to make its results more accurate, it will have to resort increasingly to a coherent implementation of natural language semantics. In this paper, we will focus on the extraction of semantic case roles from texts. After setting the essential theoretical framework, we will argue that it is possible to detect case roles on the basis of morphosyntactic and lexical surface phenomena. We will give a concise overview of our methodology and of a preliminary test that seems to confirm our hypotheses.

Introduction

Information extraction (IE) from texts currently receives a large research interest. Traditionally, it has been associated with the – often verbatim – extraction of domain-specific information from free text (Riloff & Lorenzen 1999). Input documents are scanned for very specific relevant information elements on a particular topic, which are used to fill out empty slots in a predefined frame. Other types of systems try to acquire this knowledge automatically by detecting reoccurring lexical and syntactic information from manually annotated example texts (e.g. Soderland 1999).

Most of these techniques are inherently limited because they exclude natural language semantics as much as possible. This is understandable for reasons of efficiency and genericity but it restricts the algorithms' possibilities and it disregards the fact that – at least in free text – IE has much to do with identifying semantic roles.

In most of these systems, case role detection as a goal in itself has been treated in a rather trivial way. Our research will try to provide a systematic approach to case role detection as an independent extraction task. Using notions from systemic-functional grammar and presupposing a possible mapping between morphosyntactic properties and functional role patterns, we will develop a general model for case role extraction. The idea is to learn domain-independent case role patterns from a tagged corpus, which are then (automatically) specialized to particular domain-dependent case role sets and which can be reassigned to previously unseen text. In this paper, we will focus on the first part of this task.

For IE, an accurate and speedy detection of functional case roles is of major importance, since they describe events (or states) and participants to these events and thus allow for identifying real-world entities, their properties and interactions between them.

1 Theoretical setting

One of the earliest and most notable accounts on case roles is without any doubt Charles Fillmore's groundbreaking article (Fillmore 1968). His most fundamental argument is that the notion of case is not so much connected to morphosyntactic surface realisation as to syntactico-semantic categories in the deep structure of a language. Particular constellations of case roles determine distinctive functional patterns, a considerable part of which (according to Fillmore) is likely to be universally valid. This deep-structure case system can be realized in the surface structure by means of a set of language-dependent transformation rules (see Fillmore 1968). As a consequence there has to be a regular mapping between the case system

and its surface realization – which includes case markers, word order, grammatical roles, etc. For our research, we will disregard the transformational dimension in Fillmore's theory but we will nevertheless assume that there is at least some degree of correspondence between the case role system underlying a language and its (1) morphosyntax, (2) relative word order and (3) lexicon.

In Halliday's systemic-functional grammar (Halliday 1994; Halliday & Matthiessen 1999), functional patterns that are part of the language's deep structure are organized as *figures*, i.e. configurations of case roles which consist of:

1. A nuclear process, which is typically realized by a verb phrase. Processes express an event or state as it is distinctly perceived by the language user.
2. A limited number of participants, which are inherent to the process and are typically realized by noun phrases. They represent entities or abstractions that participate in the process.
3. An in theory unlimited number of circumstantial elements. Circumstances are in most cases optional and are typically realized by prepositional or adverbial phrases. They allocate the process and its participants in a temporal, spatial, causal, … context.

Processes are classified into types and subtypes, each having its particular participant combinations. We discern four main process types: Material, Mental, Verbal and Relational (Halliday 1994). Figure 1 is an example of a verbal process, the Sayer being the participant 'doing' the process and the Receiver the one to whom the (implicit) verbal message is directed.

Invesco in merger	talks	with AIM Management
Sayer	*Verbal Process*	*Receiver*

Figure 1 – Example of a verbal process

Since these main types (and some secondary ones) correspond to universal experiential modi, it is to be expected that they will have a certain universal validity, i.e. that they are in some way or another present in all languages of the world.

For our preliminary experiments, we use a reduced version of the case role model proposed by Halliday (1994, p. 106-175), as it is a consistent, well-developed and relatively simple system, which makes it very suitable for testing the validity of our assumptions. For actual applications, we will replace it by a more elaborate variant, most likely Bateman's Generalized Upper Model (Bateman 1990; Bateman et al. in progress). Bateman's model is finer-grained than Halliday's; it is to a large extent language-independent; and it has been specifically developed for implementation into NLP systems (see Bateman et al. in progress).

2 Our approach

Given the framework outlined above, we consider case role detection to be a standard classification task. In pattern classification one attempts to learn certain patterns or rules from classified examples and to use them for classifying previously unseen instances (Hand 1997). In our case, a class is a concatenation of case roles that constitute one particular process (i.e. the deep structure figure) and the pattern itself is to be derived from the morphosyntactic and lexical properties corresponding to that process (its surface realisation).

Taking that point-of-view, individual realisations of figures – roughly corresponding to stripped-down clauses – are translated into fixed-length sets of lexical and morphosyntactic features (word order is implicitly encoded) and a functional classification is manually assigned to them. For each verb the classification algorithm then attempts to match all functional patterns to one or a few relevant sets of distinctive features. The latter are translated into patterns that can be used to match an occurrence in a text to a particular constellation of case roles.

The entire learning process consists of five main steps:

1. Preprocessing
2. Annotation
3. Feature selection
4. Training of the classifier
5. Translation into rules

In the preprocessing phase, the input text is tagged, lemmatised and chunked. The output is standardized and passed to the annotation tool,

in which the user is asked to assign case role patterns to individual clauses. For now, we will only take into account processes, participants and circumstantial elements of Extent and Location.

In a next step, individual training examples – each example corresponding to one figure – are converted to a fixed-length feature vector. For each phrase, the lexical and morphosyntactic features of the head and of the left and right context boundaries (i.e. the first and the last and each of its relevant features has a fixed position. We expect this vector representation to be relevant in most languages apart from free word order languages. Currently, our model focuses on English.

In the fourth step, the classifier is trained to discriminate features that are distinctive for each process type associated with a particular verb. These features are again translated into rules that can be used for reassigning case roles that have been learned to previously unseen text.

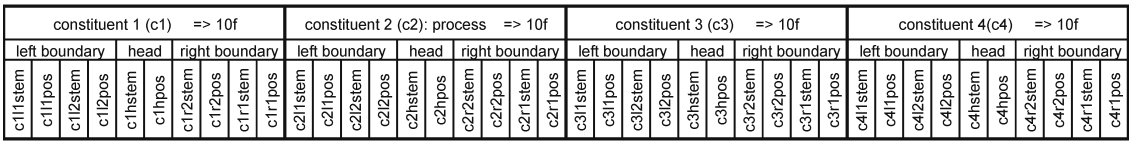

Figure 2 – The feature set

token of the strings pre- and postmodifying the head) are automatically extracted from the tagged text and added to the vector. This enables us to align corresponding features quite accurately without having to resort to any complex form of phrasal analysis. Although this reduction of the context of the head word may seem to be counter-intuitive from a grammatical point-of-view, our initial tests indicate that it does capture most constructions that are relevant to the extraction task.

Feature selection is necessary for two main reasons. Firstly, it is impossible to take into account all lexical and morphosyntactic features, since that would boost the time-complexity, incorporate many irrelevant features and bring down accuracy when a limited set of training examples is available. Secondly, natural language utterances have the uncanny habit of being of variable length. The latter aspect is problematic not only because classification algorithms usually expect a clearly delineated set of features, but also because it is crucial to align examples in order to compare correspondent features.

In our test setting, we will constrain the maximal number of case roles per figure to four. Since each case role is transformed into a set of 10 features, a figure will be translated into a 40-dimensional feature vector (see Figure 2).

As a result, a particular constellation of case roles is treated as one pattern in which each role This is necessary because the variable length of figures and – within figures – of phrases is bound to cause difficulties when applying the patterns that were learned to new sentences. Rules have the advantage over feature vectors in that they allow us to use *head-centred stretching*: when figures are assigned to previously unseen sentences and no pattern can immediately be matched, the nearest equivalent according to the head of the figure will be assigned; the rest of the pattern will be allocated by shifting the left and right context of the head towards the left and right sentence boundaries. A similar approach will be used for matching individual roles to phrases.

3 An experiment

Before engaging in the laboursome task of building a set of tools and tagging an entire corpus, we decided to test the practical validity of our ideas on a small scale on the verb *be*. We manually constructed a limited set of training examples (76 occurrences) from the new Reuters corpus (CD-rom, *Reuters Corpus. Volume 1: English Language, 1996-08-20 to 1997-08-19*) and processed it with the C4.5 classification algorithm (Quinlan 1993).

Figure 3 gives an overview of the process. The tagged text[1] (step 1) is translated into a set of

[1] For our first experiment we used TnT (http://www.coli.uni-sb.de/~thorsten/tnt/). In our

features (step 2). A functional pattern is used as the class corresponding to the feature set (the last entry in step 2). The classifier extracts one or more distinctive features (step 3), which are in turn transposed into a rule (step 4) that is used in case role assignment.

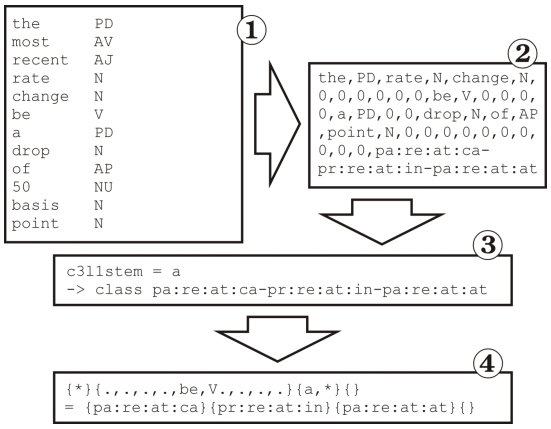

Figure 3 – Schematic illustration of the experiment

Initial results are encouraging. The evaluation component of C4.5 revealed an error rate of 9.2% when reapplying its rule extractions on the training data. Given the limited amount of data, these results are reasonable. Manual application of the rules (from step 4) to new text confirmed their natural look-and-feel. We are currently testing the approach with larger amounts of training and testing data. Most of the current errors are caused by the limited amount of training data in our experiment: in a number of cases there was only one instance of a particular figure.

4 Discussion and future improvements

Although most shortcomings that arose in our present set-up can be settled relatively easily, a number of issues still remains to be resolved. From a theoretical angle, the most urgent problem is the underspecification of the material domain (or the disagreement on exactly how material processes ought to be subclassified). Unfortunately, most verb meanings are material and distinctions in the material domain tend to be rather crucial in most IE applications.

Two major implementational difficulties are related to circumstantial elements. Since circumstances are normally not inherent to the process, they do not tend to have a fixed position in a figure. In addition, no formal parameters exist to distinguish obligatory circumstances from optional ones. Since it would be absurd to encode all variation in separate patterns, it is tempting just to add empty slots at the most predictable positions where circumstances could appear, but that would still not tell apart optional and obligatory circumstances and it would be a rather ad hoc solution. We are currently investigating whether both problems might be dealt with by encoding the relative position of case roles explicitly.

In our current information society, it will become increasingly important to extract information on well-specified events or entities from documents. Case role detection will provide a way to do this by integrating 'real' semantics into the systems without overburdening the algorithms. For instance, in our example analysis (Figure 1) we can immediately identify two entities involved in a communicative action, one that does the talking ('Invesco') and one that is being talked to ('AIM management'). An immediate application of case role detection is straightforward IE, which typically attempts to extract specific information from a text. However, the algorithm could also be used for optimising information retrieval applications, in the construction of knowledge bases, in questioning-answering systems or in case-based reasoning. Actually, for real natural language understanding a highly accurate model for interpreting case roles in some form will be unavoidable.

A major advantage of our approach is that the pattern base resulting from it will contain semantic information and yet be fully domain-independent. In a next stage of our research, we will try to specialize the generic case roles automatically to domain-dependent ones. At first sight, this two-step approach might appear cumbersome, but it will enable us to easily expand the pattern base while reusing the hard-won patterns.

---

present experiment, it is replaced by LT POS (http://www.ltg.ed.ac.uk/~mikheev/software.html).

We manually lemmatised the tokens, but we are currently using a lemmatizer based on WordNet.

5 Related research

Historically, case role detection has its roots in frame-based approaches to IE (e.g. Schank & Abelson 1977). The main problem here is that to build case frames one needs prior knowledge on which information exactly one wants to extract.
In recent years, different solutions have been offered to automatically generate those frames from annotated examples (e.g. Riloff & Schmelzenbach 1998, Soderland 1999) or by using added knowledge (e.g. Harabagiu & Maiorano 2000). Many of those approaches were very successful but most of them have a tendency to blend syntactic and semantic concepts and they still have to be trained on individual domains.
Some very interesting research on case frame detection has been done by Gildea (Gildea 2000, Gildea 2001). He uses statistical methods to learn case frames from parsed examples from FrameNet (Johnson et al. 2001).

Conclusion

There is a definite need for case role analysis in IE and in natural language processing in general. In this article, we have tried to argue that generic case role detection is possible by using shallow text analysis methods. We outlined our functional framework and presented a model that considers case role pattern extraction to be a standard classification task. Our main focus for the near future will be on automating as many aspects of the annotation process as possible and on the construction of the case role assignment algorithm. In these tasks, the emphasis will be on genericity and reusability.

Acknowledgements

We would like to thank the *Institute for the Promotion of Innovation by Science and Technology in Flanders* (*IWT-Flanders*) for the research funding.

References

Bateman J. (1990) *Upper Modelling: a general organization of knowledge for natural language processing*. In "Proceedings of the International Language Generation Workshop", Pittsburgh, June 1990.

Bateman, J. (in progress) *The Generalized Upper Model 2.0*. http://www.darmstadt.gmd.de/publish/komet/gen-um/newUM.html. Checked 15 February 2002.

Fillmore Ch. (1968) *The case for case*. In "Universals in Linguistic Theory", E. Bach & R.T. Harms, ed., Holt, Rinehart and Winston, New York, pp. 1-88.

Gildea D. (2000) *Automatic labeling of semantic roles*. Qualifying exam proposal, University of California, January 2000, 21 p.

Gildea D. (2001) *Statistical Language Understanding Using Frame Semantics*. PhD. dissertation, University of California at Berkeley, 2001, 109 p.

Halliday M.A.K. (1994) *An introduction to functional grammar*. Arnold, London, 434 p.

Halliday M.A.K. and Matthiessen C. (1999) *Construing Experience Through Meaning. A Language-Based Approach to Cognition*. Cassell, London, 657 p.

Hand D. (1997) *Construction and Assessment of Classification Rules*. Chichester: John Wiley & Sons, Chichester, 214 p.

Harabagiu S. and Maiorano S. (2000) *Acquisition of linguistic patterns for knowledge-based information extraction*. In "Proceedings of LREC-2000", Athens, June 2000.

Johnson C. et al. (2001). *The FrameNet Project: Tools for Lexicon Building*. http://www.icsi.berkeley.edu/~framenet/book.pdf. Checked 15 February 2002.

Quinlan J. (1993) *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, 302 p.

Riloff E. and Lorenzen J. (1999) *Extraction-based text categorization: generating domain-specific role relationships automatically*. In "Natural Language Information Retrieval", T. Strzalkowski, ed., Kluwer Academic Publishers, Dordrecht, pp. 167-195.

Riloff E. and Schelzenbach M. (1998) *An empirical approach to conceptual case frame acquisition*. In "Proceedings of the Sixth Workshop on Very large Corpora", Montreal, Canada, August 1998.

Schank R. and Abelson R. (1977) *Scripts, Plans, Goals and Understanding. An Inquiry into Human Knowledge Structures*. Erlbaum, Hillsdale, NJ, 248p.

Soderland S. (1999) Learning information extraction rules for semi-structured and free text. In *Machine Learning 34*, 1/3, pp. 233-272.