Minimal Recursion Semantics - Overview and Applications

term paper for the course
Die Grundlagen der Quantifikation
(Prof. Dr. Tibor Kiss)
winter term 2006 / 2007

Björn Wilmsmann
Department of Linguistics
Ruhr-University, Bochum, Germany

May 1, 2007

Contents

Abstract

While conventional semantic theory is suitable for describing most of the phenomena exposed by natural language, it also sports some serious drawbacks.

First of all, some more intricate linguistic phenomena like extraposed relative clauses ([Kiss, 2005, p. 281]) and dislocated phrases in general ([Kiss, 2002, p. 109]) impose difficulties on the most widely used implementations of semantic theory.

Secondly, while providing a high standard of expressive adequacy and compatibility with the usual grammar theories, many approaches to semantic theory prove to be computationally intractable, thus excluding them for computational applications right from the start. Which is more, some natural language processing (NLP) applications like parsing for example require underspecified semantic representations as well.

Minimal Recursion Semantics (MRS) is a fairly recent approach designed to deal with the issues mentioned above. While not establishing a theory of semantics by itself it provides a framework that is capable of, amongst other aspects, accommodating the semantics of dislocated phrases as well as being used as a fundament for NLP applications.

keywords:

- minimal recursion semantics, dislocated phrases, information extraction, ontology generation

1 Introduction

Minimal Recursion Semantics is a semantic framework designed with the following criteria for large-scale computational applications of semantic theory in mind ([Copestake et al., 2005, p. 281]):

- Expressive adequacy
- Grammatical Compatibility
- Computational Tractability
- Underspecifiability

The first two aspects are requirements of a mainly linguistic dimension. A framework meeting the requirement of expressive adequacy is capable of expressing linguistic meanings correctly. Grammatical compatibility obviously is about how the framework ties in with already existing grammars like Head-Driven Phrase Structure Grammar (HPSG) (Pollard and Sag [1994]).

The latter two requirements are clearly motivated by the necessities imposed by practical applications. Computational tractability is all about in how far it is possible to process a given structure or run a specific algorithm on that structure efficiently in a reasonable amount of time. However, this aspect is not only justified by the limits of current software technology and hardware equipment but by the limited processing capacity of the human brain as well, since a representation which cannot be resolved by a machine in finite time would also be beyond the human language faculty.

Underspecifiability finally means the ability to leave the specifics of partial representations open while still retaining the validity of the underspecified expression. This is particularly useful in chunk parsing settings for instance, because shallow parsers do not provide sufficient information for completely populating nested semantic representations ([Copestake, 2004, 2006, p. 1]).

Copestake [2004, 2006] describes three major problems limiting 'the practical utility of systems based on deep processing of text', the first one being the vast search space opening up when processing textual data with full-blown parsers delving deeply into the underlying linguistics structure. The second problem is that of deep parsing algorithms failing to parse data that is considered ungrammatical for one reason or another, the most frequent reason being insufficient lexical information due to data sparsity. The third and final problem for deep parsing systems simply is speed. Processing large amounts of texts as is required in terms of applications in information retrieval and information extraction is not feasible using this techniques, for they are way too slow.

In this work I shall outline an overview of the formalism behind Minimal Recursion Semantics and some refinements to the original framework by drawing upon the works of Copestake et al. [2005] and Copestake [2004, 2006].

Afterwards, I shall deal with some intricate linguistic phenomena which can be accommodated more easily in a comprehensive theory of grammar including semantics by using MRS (Kiss [2005] and Kiss [2002]).

Finally, I shall present some applications of the Minimal Recursion Semantics framework in information technology. These include question answering machines, information extraction systems and combined deep and shallow parsers. A very promising method for deriving ontologies from text corpora (Wikipedia in this case) using MRS will be introduced as well (Herbelot and Copestake [2006]).

2 MRS - A Formal Overview

2.1 Basic Assumptions

Minimal Recursion Semantics does not constitute a theory of semantics by itself but can rather be considered 'a meta-level language for describing semantic structures in some underlying object language' ([Copestake et al., 2005, p. 283]).

The basis structural units in MRS are so-called elementary predications (EP). An elementary predication is a single semantic relation (usually representing a single lexeme) with its required arguments, e.g. *love(x, y)*. These EPs are not licensed to be embedded in one another, which entails syntactically flat structures ([Copestake et al., 2005, p. 283]).

As is outlined in [Copestake et al., 2005, p. 284] flat structures are highly desirable for semantic represensations for various reasons:

First, syntactically recursive structures can contain information that is irrelevant in terms of semantic representations and hence imposes a considerable overhead. Consider the following alternative representations of the phrase *tiny wooden box*:

- (1) λx (tiny(x) $\wedge$ (wooden(x) $\wedge$ box(x)))
- (2) λx ((tiny(x) $\wedge$ wooden(x)) $\wedge$ box(x))

Both representations have the same meaning with just a subtle structural difference brought about by the binary nature of the $\wedge$ operator: In the first case we have a wooden box that is tiny and the second we are dealing with a box that happens to be both tiny and wooden. This difference can however be neglected in terms of truth conditional semantics, since the $\wedge$ operation is associative and bracketing therefore does not matter.

Furthermore, representing syntactic information in semantic structures can give rise to problems regarding the mapping of meaning between different languages. The semantics of an expression should be the same whether this expression be uttered in, for instance English or German, regardless of the differences in syntactic structure between those languages. For example, an expression like *white horse* is conventionally translated into German as *Schimmel*. When considering the example given above, it becomes obvious that a phrase like *white English horse* leads to a mapping problem caused by ambiguous bracketing ([Copestake et al., 2005, p. 284]).

In MRS these issues are being dealt with by reducing semantic structures to flat representations like the following ones, essentially by allowing the logical AND operator to link more than two arguments:

- (3) λx (tiny(x), wooden(x), box(x))
- (4) λx (white(x), English(x), horse(x))

However, semantic representation still needs a minimum level of recursion (hence the name MRS) in order to be able to correctly interpret scopal structures ([Copestake et al., 2005, p. 286]):

- (5) λx (every(x), horse(x), old(x), white(x))

This structure could equally mean *every horse is old and white*, *every white horse is old* or *every old horse is white*. In order to avoid such ambiguity the MRS framework at this point allows for embedded, recursive structures. As we shall see later, and tying in with the underspecifiability requirement, MRS allows scopal structures to be underspecified if either there is insufficient information for populating the structure or fully specifying the structure is not desired.

Having outlined the motivation and the basic assumptions behind MRS, in the next section I shall sketch how these aspects are formally represented in the MRS framework.

2.2 Formal Representation

The MRS framework takes several measures to reduce the complexity of semantic structures in order to achieve an almost completely flat and potentially underspecified representation ([Copestake et al., 2005, p. 287-291]).

First of all, the logical operators ∨ and ∧ are altered in that their binary structure is extended to an n-ary structure. For instance, both of these structures repeated from above

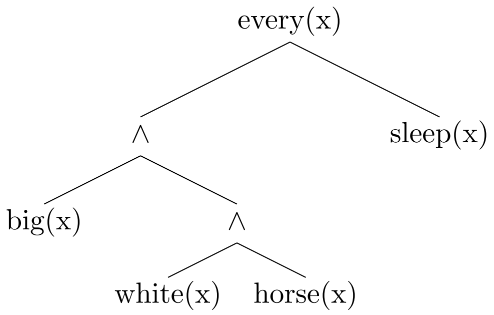
- (6) λx (tiny(x) ∧ (wooden(x) ∧ box(x)))
- (7) λx ((tiny(x) ∧ wooden(x)) ∧ box(x))

will be represented as the following group of EPs

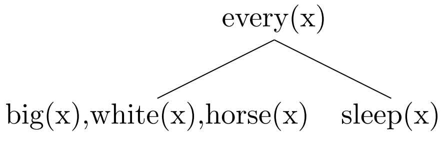
- (8) tiny(x), wooden(x), box(x)

This basically equates to merging a tree structure of depth n into a tree of depth 1 ([Copestake et al., 2005, p. 287-288]):

The deep structure



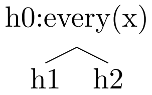
is merged into the minimally recursive structure



Please note that the $\wedge$ operator is implicit in this representation and that the group structurally is a bag rather than a set, because elements do not need to be ordered and may be repeated as well.

In order to allow for flexible and minimal recursion, in a further step the links of the tree structure underlying each non-minimally recursive representation are replaced by tags (so-called *handles*) which identify an EP with a potential scopal argument position.

In terms of data structures this means creating a tree with handles as nodes and dynamically linking the list of elements associated with each node when required ([Copestake et al., 2005, p. 289]):



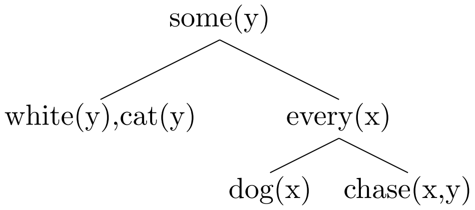
h1 = (big(x),white(x),horse(x)) h2 = (sleeps(x))

This allows us to keep structures flexible enough for accommodating scope while still retaining a minimal structure.

While the current specifications already allow us to define minimal structures, a further step has to be taken in order to meet the underspecifiability requirement. Consider the following example, taken from [Copestake et al., 2005, p. 290]:

(9) Every dog chases some white cat

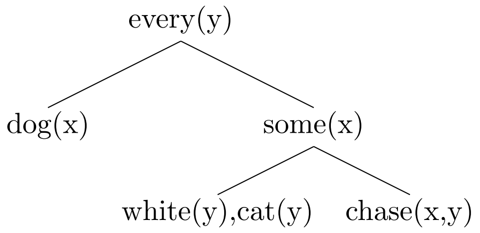
Given the formalism introduced above, this sentence, which is ambiguous due to the either wide or narrow scope of the quantifiers *every* and *some* ([Heim and Kratzer, 1998, p. 178-179]), can be interpreted in two different manners. First let us consider the one in which *some* outscopes *every*:



This tree corresponds to the following MRS structure:

h1: every(x, h3, h4), h3: dog(x), h7: white(y), h7: cat(y),
h5: some(y, h7, h1), h4: chase(x, y)

The second interpretation, in which *every* outscopes *some* can be represented as follows:



Now this second tree can be represented by using the following MRS structure:

> h1: every(x, h3, h5), h3: dog(x), h7: white(y), h7: cat(y),
> h5: some(y, h7, h4), h4: chase(x, y)

The MRS structures merely differ regarding the handles occupying the second argument position of the quantifier predicates, which is why these structures lend themselves to a further unification ([Copestake et al., 2005, p. 291]):

> h1: every(x, h3, h8), h3: dog(x), h7: white(y), h7: cat(y),
> h5: some(y, h7, h9), h4: chase(x, y)

This structure differs from the ones above in that the specific handles h4 (denoting chase(x, y)), h1 and h5 (denoting the quantifier predicates for *every* and *some*) have been replaced by two underspecified handles h8 and h9.

The structures listed above now can easily be rebuilt by either linking h8 to h5 and h9 to h4 (*every* outscoping *some*) or h8 to h4 and h9 to h1 (*some* outscoping *every*).

This finally provides us with a means of keeping semantic structures underspecified, but at the same time retaining the possibility for a later specification.

In the next sections I shall present some applications for the framework outlined in this section, starting off with how dislocated phrases like extraposed relative clauses can be accommodated in a comprehensive theory of syntax and semantics by using underspecified semantic representations.

Afterwards, I shall turn to computational applications of the MRS framework in terms of information retrieval and information extraction systems.

3 Semantic Representation of Dislocated Phrases

Dislocated phrases comprise a variety of different linguistic phenomena: Topicalization, wh-movement and extraposed relative clauses ([Kiss, 2002, p. 109])

All of these phenomena share a common feature, that is they all involve relating an element to another, non-local element. For the purpose of this section I shall focus on extraposed relative clauses (defining ones for that matter) as examples of dislocated phrases. Please consider the following sentences ([Kiss, 2005, p. 281]):

- (10) Mary mentioned **the claim that John is intelligent** yesterday.
- (11) Mary mentioned **the claim** yesterday **that John is intelligent**.

While in (10) the relative clause *that John is intelligent* is adjacent to the element it modifies (*the claim*), in (11) the modifier is not directly adjacent to the modified element, hence making the relation between the modifier and the modified element a non-local one.

Using conventional semantc represention (10) and (11) could roughly be interpreted as follows:

(12) λx (claim(x) $\wedge$ (John(x) $\wedge$ intelligent(x)))

The problem with this kind of representation is that it does not have a clear-cut boundary to the syntactic structure of the expression.

Moreover, it can only be used in cases where the expression it represents occurs as a whole. However, in cases like (11) where the elements do not appear adjacent to each other, this kind of semantic representation cannot be used, since it either fits the complete phrase or not at all.

A corresponding MRS representation helps in amending this:

(13) h1: claim(x, h2), h2: John(x), h2: intelligent(x)

This representation constructs the underlying meaning simply by intersecting the components *claim*, *John* and *intelligent* using identical handles for all entities in the intersection.

In other words the utterance is only true if it is the case that:

- 1.) a claim has been made $\wedge$
- 2.) someone has been claimed to be intelligent $\wedge$
- 3.) this someone is John

MRS thus provides us with a flexible way of representing the semantics of an extraposed relative clause and dislocated phrases in general by glossing over the syntactic variations it might display and the pragmatic consequences (like topicalization for example) these might have. In the examples above, for instance *yesterday* is emphasised in (10), whereas the relative clause *that John is intelligent* is put emphasis on in (11), both of which do not affect the core semantics of the clause.

4 Computational Applications

While the previous sections gave an overview about the fundamentals of MRS and how linguistic theory might benefit from the MRS framework, we shall now turn to computational applications of MRS.

MRS has come to be used in areas where knowledge about semantic relations between entities is important. These areas include question answering and information extraction systems in general as well as parsing applications.

Moreover, MRS has been used in a recent approach for automatically inducing ontologies, that is complex networks of semantic relations, from natural language texts (Herbelot and Copestake [2006]).

In the following section I shall outline the use of and possible benefits from the MRS framework in each of these areas.

4.1 Question Answering Systems

Question answering systems / engines (see http://www.brainboost.com/ for instance) are a special kind of combined information retrieval and information extraction systems.

They take natural language questions, as opposed to the keyword-centric syntax of common-place search engines, try to interprete the meaning the user wanted to convey and mine for excerpts of a corpus that might be an appropriate answer to the user's question - again as opposed to the results of search engines, which simply render the documents that fit the user-given keywords best.

This task requires not only a robust syntactic parser but a solid understanding of the semantics behind natural language expressions as well.

MRS lends itself to be used in question answering systems because of its underspecifiability feature. While there has to be a thorough understanding of the meaning of the question provided by the user (using deep parsing), computing the complete meaning of all potential answers in the corpus for sifting out the most appropriate ones is inefficient at best if not computationally infeasible right away. MRS can help resolve this issue by leaving semantic relations in the potential answers underspecified to a certain degree (by only performing a shallow parse), thus vastly reducing search space ([Copestake, 2004, 2006, p.2, 14]).

Once the search space is narrowed down to a small number of possible answers, deep parsing and completion of the underspecified structures could be conducted for these as well.

4.2 Information Extraction Systems

A closely related but more general field of application for MRS are information extraction systems.

While search engines only provide documents the user might find interesting, more sophisticated information extraction systems supply the user with the specific chunks of information he or she has asked for.

Apart from sub-tasks like tokenization, normalization, sentence boundary detection and part-of-speech (POS) tagging, an automatic understanding of not only the shallow morphological and syntactic structure, as provided by POS tagging, but also about the semantic relations between potentially useful chunks of information might come in handy.

However, when processing large corpora (as information extraction systems usually do), we face the same issue as with question answering systems: Deep-parsing all sentences in a large corpus is computationally impossible. Hence, an approach like MRS which allows us to underspecify relations might help in finding the presumably most interesting sentences (those with the highest potential for containing the information sought for), whose structure then can be parsed completely in order to retrieve the requested chunk of information ([Copestake, 2004, 2006, p.2, 14]).

4.3 Deep and Shallow Parsers

Another way of using MRS I would like to come up with is integrating deep and shallow parsing systems.

A possible problem for deep parsing systems is that they require specific lexical and grammatical information as an input. If the lexical information for the words in a sentence, their respective parts of speech or the way in which they are combined are not known to the system, deep parsing is bound to fail.

This is where shallow parsers with less constraints than a deep parser might be used first in order to enrich the input and produce a structure that is more readily understood by the deep parser.

MRS is quite suitable for this task since a shallow parse iteration could simply provide an underspecified MRS structure, which will be understood and specified in a deep parse iteration ([Copestake, 2004, 2006, p.2, 15]).

4.4 Ontology Induction using MRS

In this section I shall give a brief overview of a recent approach for automatically acquiring ontologies from natural language corpora (Herbelot and Copestake [2006]) by using Robust Minimal Recursion Semantics (RMRS) as introduced in Copestake [2004, 2006].

Ontologies and the Semantic Web (Berners-Lee et al. [2001]) have been proven to be very useful in a variety of areas that require computers to have an understanding of relations between entities. However, bootstrapping, that is creating ontologies for use with Semantic Web techniques in the first place, usually still requires a tedious manual process of sifting through potential entities and the relations between them.

In order to alleviate this issue, several approaches have been suggested (see Mani et al. [2004] for example). One such approach is the one introduced in Herbelot and Copestake [2006]. In this article, Herbelot and Copestake rely on the MRS framework for finding relations between entities. They point out that while previous approaches 'have focused on generalised is-a or part-of relationships' ([Herbelot and Copestake, 2006, p.2]) their framework is capable of 'extracting general hyponymic relations'([Herbelot and Copestake, 2006, p.2]).

During their work the RMRS system has been used to derive taxonomic relationships between entities from 12.000 animal-related biological articles taken from the online encylopaedia Wikipedia ([Herbelot and Copestake, 2006, p.1-2]).

Initial experiments led to an extraction of 4771 taxonomic relations, yielding a rough recall of 20 % and a precision of 88.5 % ([Herbelot and Copestake, 2006, p.10]).

The authors consider these results to be sufficiently promising to venture further in this direction ([Herbelot and Copestake, 2006, p.10]):

> 'These promising initial results call for the development of a more complete system, including pattern extraction and pattern matching features.'

The issues the authors would like to deal with in this more complete system are mostly caused by insufficient normalisation (stemming, singular <-> plural conversion and compound word order) and data sparsity ([Herbelot and Copestake, 2006, p.7]).

5 Conclusion

Minimal Recursion Semantics (MRS) is a framework used for representing semantic relations in an accurate fashion that is at the same time compatible with other grammar frameworks, computationally tractable and last not least allows for underspecification of semantic structure (Copestake et al. [2005] and Copestake [2004, 2006]).

I have presented an overview of the most important aspects of MRS and the formalism in which these are realised. I have tried to underline the advantages the MRS framework by comparing it to the formalism of conventional semantics.

MRS sports several features that make it especially suitable in terms of some linguistic phenomena and computational applications of semantic theory alike (Copestake et al. [2005]):

- it expresses semantic relations adequately
- it is compatible to other grammar frameworks
- it is computationally tractable
- MRS structures are underspecifiable

After having introduced the framework itself, we have turned to applications both in linguistic theory and in various kinds of information extraction systems which at least to some extent rely on knowledge about the underlying semantics of texts for accomplishing their respective task.

First of all, I have briefly discussed extraposed relative clauses (and dislocated phrases in general) for motivating the usage of Minimal Recursion Semantics in linguistic theory. It could be shown that MRS proves to be highly useful whenever it comes to interpreting the meaning of phrases whose constituents are not directly adjacent to each other (Kiss [2005] and Kiss [2002]).

Furthermore, I have introduced several practical applications which could possibly profit from using the MRS framework. These applications included automatic question answering, information extraction and combined deep and shallow parsing (Copestake [2004, 2006]).

Finally, I have presented an approach for acquiring ontological relationships from the online encyclopaedia Wikipedia by using the Robust Minimal Recursion Semantics (RMRS) system (Herbelot and Copestake [2006]). Yielding a rough recall of 20 % and a precision of 88.5 % ([Herbelot and Copestake, 2006, p.10]) in the initial version, this approach appears to be promising as well.

Putting all this into consideration, we can say that Minimal Recursion Semantics is a flexible framework for expressing semantic relations that is equally useful in linguistic theory as in practially applications which make use of this theoretical background.

An area that might lend itself to further research is the acquisition of ontologies not only from corpora which, like biological articles, contain distinct taxonomic relationships, but from corpora from different areas, like engineering for instance, which might not only expose simple hyponymic relations but more complex ones, too.

References

T. Berners-Lee, J. Hendler, and O. Lassila. The Semantic Web. Online version only, 2001. URL `http://www.sciam.com/print_version.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21`.

A. Copestake. Robust Minimal Recursion Semantics. Online version only, 2004, 2006. URL `http://www.cl.cam.ac.uk/~aac10/papers/rmrsdraft.pdf`.

A. Copestake, D. Flickinger, C. Pollard, and I. Sag. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3 (4):281–332, December 2005. URL `http://lingo.stanford.edu/sag/papers/copestake.pdf`.

I. Heim and A. Kratzer. *Semantics in Generative Grammar*. Blackwell textbooks in linguistics. Blackwell, Malden, Massachusetts, 1998.

A. Herbelot and A. Copestake. Acquiring Ontological Relationships from Wikipedia Using RMRS. In *Proc.of the ISWC 2006 Workshop on Web Content Mining with Human Language Technologies*, 2006. URL `http://orestes.ii.uam.es/workshop/12.pdf`.

T. Kiss. Semantic Constraints on Relative Clause Extraposition. *Natural Language and Linguistic Theory*, 23(2):281–334, May 2005. ISSN 0167-806X. URL `http://www.linguistics.ruhr-uni-bochum.de/~kiss/publications/semconst_ss.pdf`.

T. Kiss. Phrasal typology and the interaction of topicalization, wh-movement and extraposition. In *Proceedings of the 9th International Conference on Head-Driven Phrase Structure Grammar*, pages 109–128. CSLI Publications, 2002. URL `http://www.linguistics.ruhr-uni-bochum.de/~kiss/publications/hpsg02.pdf`.

I. Mani, K. Samuel, K. Concepcion, and D. Vogel. Automatically inducing ontologies from corpora. In S. Ananadiou and P. Zweigenbaum, editors, *COLING 2004 CompuTerm 2004: 3rd International Workshop on Computational Terminology*, pages 47–54, Geneva, Switzerland, August 29 2004. COLING. URL `http://complingone.georgetown.edu/~prot/data/coling-workshop.pdf`.

C. Pollard and I. A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, Chicago, Illinois, 1994.