A Construction Grammar Method for Disambiguating Swedish Compounds

Robert Östling

Department of Linguistics
Stockholm University
SE-106 91 Stockholm
robert@ling.su.se

Abstract

This study discusses the structure of Swedish compounds within the framework of Construction Grammar, and applies the result to Word Sense Disambiguation of compound components. A construction-based approach is shown to achieve significantly better results than a set of baselines.

1. Introduction

Construction Grammar is a family of related theories of language, which hold that human language is made up of *constructions*, pre-fabricated and often idiosyncratic chunks of language which are learned and re-used. Construction Grammar theories make no strict division between syntax/morphology and semantics/pragmatics, but constructions and the way in which they are combined depend on both form and function (Goldberg, 2003; Fried and Östman, 2004).

1.1 Constructions

At a very abstract level, a *construction* is a form-meaning pair. These can be defined at different levels, informally exemplified in Table 1. While the meaning of a construction as a rule *motivates* the meaning of constructions similar in form, as is the case with the constructions [*noun*]-**s** and **eye** together motivating **eyes**, the meaning of a construction is autonomous of its parts and can show various levels of compositionality, as can be seen in the highly non-compositional idiom **catch** [*person's*] **eye**.

Form	Meaning
[*noun*]-**s**	several (of a noun)
eye	an organ of sight
eyes	several organs of sight
catch [*person's*] **eye**	to attract someone's attention

Table 1: Examples of constructions

1.2 Swedish Compounds as Constructions

Swedish has a system of productive compounding, similar to German. Apart from the theoretical interest in classifying compounds, the infinite amount of potential compounds (and the extremely large amount actually used) makes compound analysis an important practical matter in Natural Language Processing (NLP).

Construction Grammar offers a way to classify compounds as generalizations at different levels, all of which may affect the interpretation of a given compound. Take for instance the compound **plastpappa** (literally *plastic father*). Together with compounds such as **plastmormor** (*plastic grandmother*) and **plastförälder** (*plastic parent*) it is motivated by a construction of the form **plast**-[*relative*], with a meaning similar to but with different connotations than **styv**-[*relative*] (step-[*relative*].)

This construction is motivated by the more basic and common construction of the form **plast**-[*object*], which is an object made of plastic, often with connotations of being a cheap copy of something. This in turn is motivated by an even more general construction, [*material*]-[*object*], which covers a wide variety of materials and objects.

2. Disambiguation in Compounds

The first problem when analyzing a Swedish compound is where to split it, a problem that is relatively easy to address by statistical methods (Sjöbergh and Kann, 2004; Sjöbergh and Kann, 2006).

Even if we know at which positions a Swedish compound string should be split, there are two more levels of ambiguity to deal with. **Mossflora** can be unambiguously split as **moss**-**flora**, but there are two different words that both use the compound form **moss**-: *mossa* (moss, various plants) and *mosse* (moss, bog). This is ambiguity not found outside compounds. Additionally, as in English, *flora* is polysemous and can refer to either plant life or a catalog of it.

2.1 Compound Splitting with SALDO

Using the SALDO semantic and morphological database of Swedish (Borin and Forsberg, 2009), it is fairly straightforward to split a compound into its syntactic components (SALDO *lemmas*, homographs with identical inflectional paradigms). Once the set of possible lemmas that make up a compound has been determined, it is trivial to look up the semantic components (SALDO *lexemes*) that can be represented using these lemmas.

In this study I am interested in determining *which* of the lexemes is used in a given compound, a task roughly equivalent to classical Word Sense Disambiguation tasks, but where the context is the other part of the compound rather than the surrounding text.

2.2 Identifying Constructions

Using a catalog of Swedish compound constructions, we could reduce the task of selecting the right lexeme in a compound to the task of identifying which constructions the compound is motivated by.

To find the set of constructions of the form -**kyrka** (church) in the sense of a building for worship, not in its other sense of a religious movement, we look at some of its most common instances: **stenkyrka** (stone church), **träkyrka** (wooden church), **sockenkyrka** (parish church), **stadskyrka** (city church). Within these few examples, we see two constructions: one describing churches made of a certain material, and one describing a certain area's church.

I do not currently attempt to explicitly describe these constructions, but unknown compounds of the form -**kyrka**, for instance **slottskyrka** (castle church), are compared to examples in the training data by means of graph distance in SALDO:s semantic hierarchy. In this case, the training instances **sockenkyrka** and **stadskyrka** should ensure that **slottskyrka** fits nicely into this set of constructions.

The set of constructions of the form -**kyrka** using the other sense of *kyrka* (a religious movement) would generally *not* include opposite lexemes of the same categories (building materials, administrative areas), and the different lexemes corresponding to different senses of *kyrka* can be differentiated.

One disadvantage of not using any external context is that it becomes difficult to deal with certain compounds whole form match several constructions using different senses of the components. For instance, in the compound **statskyrka** (*state church*), *state* can be used in its geographical or its (correct) political sense, and *church* in its physical or its (correct) religious sense. The form **statskyrka** could therefore fit into either of the [*geographical area*]-[*physical church*] construction or the (correct) [*political entity*]-[*religious movement*] construction.

3. Results

Table 2 summarizes an experiment where 534 compounds that are ambiguous with respect to SALDO lexemes were disambiguated using four different methods:

- **Constructions**, select the lexemes that best match the construction sets described above, or the most frequent lexeme if no matching training examples are found.
- **Most frequent**, always select the most frequent lexeme using the given lemma.
- **Coherence**, select the lexeme pair where the components are the closest to each other according to the SALDO semantic hierarchy.
- **Random**, select random lexemes.

The training data is obtained from SALDO, which contains about 9 400 compounds (most of them ambiguous) where the lexemes of the parts are given, and from a set of about 180 000 unambiguous (with respect to SALDO lemmas and lexemes) compounds extracted automatically from a corpus. Without the unambiguous compounds in the training data, recall drops to about 69%, just barely above the baseline. This is an interesting result, since there is obviously no overlap between the ambiguous compound segments being disambiguated, and the unambiguous compound segments in this training set. Rather than training a classifier to tell which sense of a segment is intended based on the opposite segment, which would require large amounts of (expensive) disambiguated compounds, we can use large amounts of (free) unambiguous compounds to find compound *constructions*.

Method	**Recall**
Constructions	389 (73%)
Most frequent	354 (66%)
Coherence	300 (56%)
Random	178 (33%)

Table 2: Results

The test data set consists of 534 ambiguous compounds disambiguated by hand (and 442 unambiguous, which are trivial and not included in the statistics).

4. Summary and Future Work

Using a method based on Swedish compound constructions, recall on a compound component disambiguation task was enhanced from 66% (the best-performing *most frequent sense* baseline) to 73%. The only data sources used are SALDO (Borin and Forsberg, 2009) and an unannotated list of compounds.

Context is an important factor in any Word Sense Disambiguation task, and methods of using further context in tandem with the methods presented here should be explored.

Furthermore, experiments are planned to explicitly extract and build a database of compound constructions in Swedish.

5. References

L. Borin and M. Forsberg. 2009. All in the family: A comparison of SALDO and WordNet. In *Proceedings of the Nodalida 2009 Workshop on WordNets and other Lexical Semantic Resources - between Lexical Semantics, Lexicography, Terminology and Formal Ontologies.*

Mirjam Fried and Jan-Ola Östman, editors. 2004. *Construction Grammar in a cross-language perspective*. John Benjamins.

Adele E. Goldberg. 2003. Constructions: a new theoretical approach to language. *TRENDS in Cognitive Sciences*, 7(5):219–224.

Jonas Sjöbergh and Viggo Kann. 2004. Finding the correct interpretation of swedish compounds, a statistical approach. In *In Proc. 4th Int. Conf. Language Resources and Evaluation (LREC*, pages 899–902.

Jonas Sjöbergh and Viggo Kann. 2006. Vad kan statistik avslöja om svenska sammansättningar? *Språk och stil*, 1:199–214.