Towards unsupervised learning of constructions from text

Krista Lagus, Oskar Kohonen and Sami Virpioja
Adaptive Informatics Research Centre
Helsinki University of Technology
P.o.Box 5400, 02015 TKK, Finland
{krista.lagus,oskar.kohonen,sami.virpioja}@tkk.fi

Abstract

Statistical learning methods offer a route for identifying linguistic constructions. Phrasal constructions are interesting both from the viewpoint of cognitive modeling and for improving NLP applications such as machine translation. In this article, an initial model structure and search algorithm for attempting to learn constructions from plain text is described. An information-theoretic optimization criteria, namely the Minimum Description Length principle, is utilized. The method is applied to a Finnish corpus consisting of stories told by children.

1 Introduction

How to represent meaning is a question that has for long stimulated research in various disciplines, including philosophy, linguistics, artificial intelligence and brain research. On a practical level, one must find engineering solutions to it in some natural language processing tasks. For example, in machine translation, the translations that the system produces should reflect the intended meaning of the original utterance as accurately as possible.

One traditional view of meaning in linguistics (exemplified e.g. by Chomsky) is that words are seen as basic blocks of meaning, that are orthogonal, i.e., each word is seen as individually conveying totally different properties from all other words (this view has been promoted e.g. by Fodor). The meaning of a sentence, on the other hand, has been viewed as compositional, i.e., consisting of the meanings of the individual words.

Idioms and other expressions that seem to violate against the principle of compositionality (e.g. *"kick the bucket"*) have been viewed as mere exceptions rather than central in language. While such a view might be convenient for formal description of language, and offers a straightforward basis for computer simulations of linguistic meaning, the view has for long been regarded as inaccurate. The problems can also observed in applications such as machine translation. Building a system that translates one word at a time yields output that is incorrect in form, and most often also its meaning cannot be understood.

A reasonable linguistic approach is offered by constructionist approaches to language, where language is viewed as consisting of *constructions*, that is form-meaning pairs.[1] The form component of the construction is not limited to a certain level of language processing as in most other theories, but can as well be a morpheme (*anti-*, *-ing*), a word, an idiom (*"kick the bucket"*), or a basic sentence construction (*SUBJ V OBJ*). The meaning of a sentence is composed from the meanings of the constructions present in the sentence. Construction Grammar is a usage-based theory and does not consider any linguistic form more basic than another. This is well aligned with using data-oriented learning approaches for building wide coverage NLP applications.

We are interested in identifying the *basic information processing principles* that are capable of producing *gradually more abstract representations* that are useful for intelligent behavior irrespective of the domain, be it language or sensory information, and irrespective of the size of the time window being analysed. There is evidence from brain research that the exactly same information-processing and learning principles are in effect in many different areas of the cortex. For example, it was found in (Newton and Sur, 2004) that if during development visual input pathways are re-routed to the region that normally contains auditory cortex, quite typical visual processing and representations ensue, but in this case in the auditory cortical area. The cortical learning al-

[1] For an overview see, e.g., Goldberg (2003).

gorithm and even the model structure can therefore be assumed identical or very similar for both processes. The differences in processing that are seen in the adult brain regions are thus largely due to each region being exposed to data with different kinds of statistical properties during individual growth.

In this article we describe our first attempt at developing a method for the discovery of constructions in an unsupervised manner from unannotated texts. Our focus is on constructions involving a sequence of words and possibly also abstract categories. For model search we apply an information-theoretic learning principle namely Minimum Description Length (MDL).

We have applied the developed method to a corpus of stories told by 1–7 year old Finnish children, in order to look at constructions utilized by children. Stories told by an individual involve entities and events that are familiar to the teller, albeit the combinations and details may sometimes be very imaginative. When spontaneously telling a story, one employs one's imagination, which in turn is likely to utilise one's entrenched representations regarding the world. Of particular interest are the abstract representations that children have—this should tell us about an intermediate stage of the development of the individual.

2 Related work on learning constructions

Constructions as form-meaning pairs would be most naturally learned in a setting where both form and meaning is present, such as when speaking to a robotic agent. Unfortunately, in practice, the meaning needed for language processing is highly abstract and cannot easily be extracted from natural data, such as video. Therefore time consuming hand-coding of meaning is needed and, consequently, the majority of computational work related to learning constructions has been done from text only. A notable exception is Chang and Gurevich (2004) who examine learning children's earliest grammatical constructions, in a rich semantic context.

While learning from text only is unrealistic as a model for child learning, such methods can utilize the large text corpora and discover structure useful in NLP applications. They illustrate that statistical regularities in language form is also involved in learning. Most work has been done within a traditional syntactic framework and thus focuses on learning context-free grammars (CFG) or regular languages. While it is theoretically possible to infer a Probabilistic Context-Free Grammar (PCFG) from text only, in practice this is largely an unsolved problem (Manning and Schütze, 1999, Ch. 11.1). More commonly, applications use a hand crafted grammar and only estimate the probabilities from data. There are some attempts at learning the grammar itself, both traditional constituent grammar and also other alternatives, such as dependency grammars (Zaanen, 2000; Klein and Manning, 2004).

Also related to learning of constructions are the methods that infer some structure from a corpus without learning a complete grammar. As an example, consider various methods that are applied to finding collocations from text. Collocations are pairs or triplets of words whose meanings are not directly predictable from the meanings of the individual words, in other words they exhibit limited compositionality. Collocations can be found automatically from text by studying the statistical dependencies of the word distributions (Manning and Schütze, 1999, Ch. 5).

Perhaps most related to construction learning is the ADIOS system (Solan et al., 2005), which does not learn explicit grammar rules, but rather generalizations in specific contexts. It utilises pseudo-graph data structures and seems to learn complex and realistic contextual patterns in a bottom-up fashion. Model complexity appears to be controlled heuristically. The method described in this paper is similar to ADIOS in the sense that we also use information-theoretic methods and learn a model that extracts highly specific contextual patterns from text. At this point our method is much simpler; in particular, it cannot learn as general patterns. On the other hand, we explicitly optimize model complexity using a theoretically well motivated approach.

3 Learning constructions with MDL

A particular example of an efficient coding principle is the Minimum Description Length (MDL) principle (Rissanen, 1989). The basic idea resembles that of Occam's razor, which states that when one wishes to model phenomenon and one has two equally accurate models (or theories), one should select the model (or theory) that is less complex. In practice, controlling model complexity is essential in order to avoid overlearning, i.e., a sit-

uation where the properties of the input data are learned so precisely that the model does not generalise well to new data.

There are different flavors of MDL. We use the earliest, namely the two-part coding scheme. The cost function to minimize consists of (1) the cost of representing the observed data in terms of the model, and (2) the cost of encoding the model. The first part penalises models that are not accurate descriptions of the data, whereas the second part penalises models that are overly complex. Coding length is calculated as the negative logarithm of probability, thus we are looking for the model $\mathcal{M}^*$:

$$\mathcal{M}^* = \arg\min_{\mathcal{M}} L(\text{corpus}|\mathcal{M}) + L(\mathcal{M}). \quad (1)$$

The two-part code expresses an optimal balance between the specificity and the generalization ability of the model. The change of cost can be calculated for each suggested modification to the model.

Earlier this kind of MDL-based approach has been applied successfully in unsupervised morphology induction. For example, the language-independent method called Morfessor (Creutz and Lagus, 2002; Creutz and Lagus, 2007) finds from untagged text corpora a segmentation for words into morphemes. The discovered morphemes have been found to perform as good as or better than linguistic morphemes or words as tokes for language models utilized in speech recognition (Creutz et al., 2007). It is therefore our hypothesis that a similar MDL-based approach might be fruitfully applied on the sentence level as well, to learn a "construction inventory" from plain text.

3.1 Model and cost function

The constructions that we learn can be of the following types:

- word sequences of different lengths, e.g., `went to`, `red car`, `and`
- sequences that contain one *category*, where a category refers simply to a group of words that is expected to be used within this sequence, i.e. `went to buy [X], [X] was`.

If only the former kind of structure is allowed, the model is equivalent to the Morfessor Baseline model(Creutz and Lagus, 2002), but for sentences consisting of words instead of words consisting of letters. Initial experiments with such a model showed that while the algorithm finds sensible structure, the constructions found are very redundant and therefore impractical and difficult to interpret. For these reasons we added the latter construction type. However, allowing only one category is merely a first approximation, and later we expect to consider also learning constructions with more than one abstract category.

The coding length can be calculated as the negative logarithm of the probability. Thus, we can work with probability distributions instead. In the likelihood we assume that each sentence in the corpus is independent and that each sentence consists of a bag-of-constructions:

$$P(\text{corpus}|\mathcal{M}) = \prod_{i}^{N} P(s_i|\mathcal{M})$$

$$P(s_i|\mathcal{M}) = \prod_{j}^{M_i} P(\omega_{ij}|\mu_{ij}, \mathcal{M}) P(\mu_{ij}|\mathcal{M})$$

where s_i denotes the i:th sentence in the corpus of N sentences, M_i is the amount of constructions in s_i, μ_{ij} denotes a construction in s_i and ω_{ij} is the word that fills the category of the construction (if the construction has a category, otherwise that probability = 1). The probabilities $P(\mu_{ij}|\mathcal{M})$ and $P(\omega_{ij}|\mu_{ij}, \mathcal{M})$ are multinomial distributions, whose parameters need to be estimated.

When using two part codes the coding of the model may in principle utilize any code that can be used to decode the model, but ideally the code should be as short as possible. The coding we use is shown in Figure 1. We apply the following principles: For bounded integer or boolean values (fields 1, 2.1, 2.3, 2.4 and 4.1 in Figure 1) we assume a uniform distribution over the possible values that the parameter can take. This yields a coding length of $\log(L)$, where L is the amount of different possible values. For the construction lexicon size (field 1), L is the number of n-grams in the corpus and its coding length is therefore constant.

When coding words (fields 2.2 and 4.2) we assume a multinomial distribution over all the words in the corpus, and the parameters are estimated from corpus frequencies. Thus the probability of construction lexicon units (field 2.2) is given by:

$$P(\text{words}(\mu_k)) = \prod_{j}^{W_k} P(w_{kj}), \quad (2)$$

1. Number of constructions	2. Constructions μ_i:	3. Construction counts	4. Categories:
	2.1. length(μ_i)		4.1. number of words
	2.2. words in μ_i		4.2. words
	2.3. has category?		4.3. word counts
	2.4. category position		

Figure 1: Coding scheme for the model.

where W_k is the number of words in construction μ_k and $P(w_{kj})$ the probability of a word. The category words (field 4.2) are coded in a similar manner.

We also need to encode the parameters for the multinomials $P(\mu_{ij}|\mathcal{M})$ and $P(\omega_{ij}|\mu_{ij}, \mathcal{M})$. We do this by encoding the corresponding counts (fields 3 and 4.3), from which the probabilities can be calculated. We use the following reasoning: If there are M different construction or word types and the sum of their counts is K, then there are $\begin{pmatrix} K-1 \\ M-1 \end{pmatrix}$ ways of choosing M positive integers so that they sum up to K. Thus the coding length is the negative logarithm of

$$P(\text{count}(\mu_1), .., \text{count}(\mu_M)) = 1/\begin{pmatrix} K-1 \\ M-1 \end{pmatrix}. \quad (3)$$

3.2 Search algorithm

Because we are optimizing both model parameters and model size at the same time, standard probabilistic parameter estimation methods, such as Expectation-Maximization, cannot be used. Instead we use an incremental algorithm for optimizing the cost function as follows: At all times we maintain a certain analysis of the corpus and try to improve it. For a given analysis it is possible to estimate the maximum likelihood parameters for $P(\omega_{ij}|\mu_{ij}, \mathcal{M})$ and $P(\mu_{ij}|\mathcal{M})$ and then calculate the cost function for that model.

The optimization proceeds with the following steps: (1) Initialize the analysis so that each word is a construction by itself and there exist no other constructions. (2) Generate all possible constructions of length ≤ 6 from the corpus. For those constructions that exist more than 10 times in the corpus, calculate the likelihood ratio. Since the likelihood side of the optimization is completely local one can calculate the change in likelihood that one would get from modeling a set of sentences using a certain construction, compared to the initial analysis. (3) In the descending order of likelihood ratios, apply the construction to all sentences where applicable. Then calculate the value of the cost function. If the change improved the cost, accept it, otherwise discard the change. Finally, proceed with the next construction.

4 Experiments

We applied our MDL-based model to a corpus consisting of stories told by Finnish children. The are several reasons for this choice of data. If one is interested in underlying cognitive processes and their development, it may be more fruitful to look at the outputs of a cognitive system in the middle of its development rather than modeling the outputs of the fully developed system. Because the data that children hear is produced by adult systems, some of it is likely to be discarded by children by means of attentional selection, and one cannot easily know which part. This problem is avoided by only looking at data that is known to be represented by the children, that is, produced by them. From the practical point of view, as we have no means of quantitative evaluation, we want to apply the method to such a data that should have many frequent and simple constructions to observe.

4.1 Corpus and preprocessing

The corpus contains 2642 stories told by children to an adult—typically a day care personnel or a parent—who has written the story down exactly as it was told, without changing or correcting anything. A minority of the stories were told together by a pair or by a group of children. The children ranged from 1 to 7 years. The story markup contains the age and the first name(s) of the storyteller(s). The stories contain a lot of spoken-language word forms. For a more extensive description of the corpus, see (Klami, 2005).

A story told by Oona, 3 years: *Mun äitin nimi on äiti. Mun iskän nimi on iskä. Iskä tuli mun kanssa tänne. Mun nimi on Oona. Jannen nimi on Janne.* A story told by Joona, 5 years and

11 months: *Dinosaurus meni kauppaan osti sieltä karkkia sitten se meni kotiin ja söi juuston. Sitten se meni lenkille ja se tappoi pupujussin iltapalaksi ja sitten se meni uudestaan kauppaan ja se ei näkenyt mitään siellä kun kauppa oli kiinni.*

The stories are preprocessed as follows: Story mark-up containing headers etc. is removed, any punctuation is replaced with a symbol # and the story is divided into sentences. After removal of story mark-up the total number of sentences in the corpus is 36,542. The number of word tokens is 244,274 and word types 24,242. Each sentence is then given as input for the construction learner.

subsectionResults

Figure 2 shows the most frequent constructions that the algorithm has discovered. One can see that the frequent constructions found by the algorithm are good, in the sense that they are not random frequent strings, but often meaningful constructions. An especially nice example of a construction found is `olipa kerran [X]`, which is the archetypical way of beginning a fairy tale in Finnish (*once upon a time there was a ...*). The prominence of `ja sitten` is caused by many stories following a pattern where the child explains some event, then uses `ja sitten` to move on to the next event and so on. The algorithm has discovered one piece of this pattern. We also see that the algorithm has discovered that the spoken language forms of `sitten` (*then*)—`sit`, `sitte` and `sitt`—are similar.

When looking at the categories, it can be seen that they are sometimes overly general. E.g., `meni metsään` and `meni #` are analysed as `meni [X]`, where in the former case `[X]` is the argument of the verb, and in the latter the verb takes no arguments, but happens to be at the end of a sentence. However, in many cases the discovered categories appear to consist of one or a few semantic or part-of-speech categories. E.g., `söi [X] #` (*ate [X] #*) contains mostly edible arguments `banaania` (*banana*), `mansikkaa` (*strawberry*), `jäniksen` (*a rabbit*) or a pronoun `hänet` (*him/her*), `ne` (*them*).

Whereas these frequent constructions are fairly good, the analyses of individual sentences generally leave much available structure unanalysed. Consider the analysed sentence: `että hirveä hai tuli niitten [perään {X → ja}] [söi {X → ne}] #` (*that terrible shark came them [after {X → and}] [ate {X → them}] #*). We can see that most of the sentence is not analysed as any abstract construction. Looking at the corpus, we can see possible constructions that the algorithm does not discover. E.g., constructions such as `[X] hai`, where the category contains adjectives or `hai [X]` where the category contains an action verb. Note also that both constructions could not currently be used at the same time, but one would have to choose either.

5 Discussion

As this is our first attempt at learning a construction inventory, there are still many things to consider. Regarding learning of the model, one a more local updating step, in addition to the current global update, would be needed. Also, the algorithm should consider merging categories that have partially overlapping words.

Currently the model structure allows only a very restricted set of possible constructions, namely exact phrases and partially filled constructions that have exactly one abstract category that can be filled by one word. It is later possible to relax both constraints, and allow a category to be filled by several consecutive words, as well as allowing many abstract categories per construction. However, adding such abstraction capability will increase the search space of possible models quite radically, bringing the complexity close to learning a PCFG from unannotated text.

Starting simple is thus prudent: we wish to ensure learnability of the model. Moreover, we wish to identify the simplest possible approach and model structure that can account for interesting and complex phenomena, when applied throughout a corpus. A possible alternative to PCFGs would be to keep the constructions simple, but allow them to overlap each other.

Our goals include also applying the found constructions to NLP applications such as machine translation. The current statistical machine translation systems solve the problems of non-compositionality by translating a longer sequence of words (phrase) at a time. However, finding the phrase pairs is usually quite heuristic, and the phrases do not include any abstract categories. Even a reasonably simple algorithm for finding more abstract constructions should help alleviate the data sparsity problems. Applying construction learning into applications is also useful as a way of evaluating the results, as there is no "gold standards" for direct automatic evaluation.

Most frequent constructions of two words		
Freq.	**Form**	**Category words (freq.)**
891	hän [X]	meni (68), oli (50), lähti (32), löysi (29), otti (19)
	he [X]	*went, was, left, found, took*
885	ja sitten	
	and then	
798	[X] on	se (82), hän (24), täällä (20), tässä (20), nyt (17)
	[X] is	*it, he/she, here, here, now*
768	meni [X]	metsään (33), ulos (33), sinne (30), # (25), nukkumaan (18)
	went [X]	*(into the) forest, outside, there, #, (to) sleep*
694	sit [X]	se (302), ne (81), kun (20), hän (17), # (12)
	then [X]	*it, they, when, he/she, #*
Most frequent constructions of three words		
632	ja [X] se	sitten (303), sit (155), sitte (109), sitt (18), kun (5)
	and [X] it	*then, then, then, then, when*
337	[X] se meni	sitten (125), sit (66), sitte (58), ja (35), kun (14)
	[X] it went	*then, then, then, and, when*
245	olipa kerran [X]	pieni (8), tyttö (7), yksi (6), koira (6), hiiri (5)
	once (upon a time) there was (a) [X]	*little, girl, one, dog, mouse*
235	ja [X] ne	sitten (129), sit (37), sitte (28), kun (6), niin (5)
	and [X] they	*then, then, then, when, so*
197	ja [X] tuli	sitten (91), se (9), sinne (6), ne (4), niistä (3)
	and [X] came	*then, it, there, they, (of) them (be-)*

Figure 2: The most frequent two- and three word constructions with their five most frequent category words.

6 Conclusions

We share the intuition found in cognitive linguistics in general, that constructions are able to capture something essential about the cognitive representations that are also the basis of our actions and situatedness in the world.

It is our hope that the study of constructions, and the endeavour of learning them from corpora and perhaps later from richer behavioral and perceptual contexts might eventually provide a new opening in the field of modeling both language and cognition.

References

N. Chang and O. Gurevich. 2004. Context-driven construction learning. In *Proc. CogSci 2004*, Chicago.

M. Creutz and K. Lagus. 2002. Unsupervised discovery of morphemes. In *Proc. Workshop on Morphological and Phonological Learning of ACL'02*, pages 21–30, Philadelphia, Pennsylvania, USA.

M. Creutz and K. Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, 4(1).

M. Creutz, T. Hirsimäki, M. Kurimo, A. Puurula, J. Pylkkönen, V. Siivola, M. Varjokallio, E. Arisoy, M. Saraçlar, and A. Stolcke. 2007. Morph-based speech recognition and modeling of out-of-vocabulary words across languages. *ACM Transactions on Speech and Language Processing*, 5(1).

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

M. Klami. 2005. Unsupervised discovery of morphs in children's stories and their use in self-organizing map -based analysis. Master's thesis, University of Helsinki, Department of General Linguistics, Helsinki, Finland.

D. Klein and C. D. Manning. 2004. Corpus-based induction of syntactic structure: Models of dependency and constituency. In *Proc. ACL 2004*, pages 478–485, Barcelona, Spain.

C. D. Manning and H. Schütze. 1999. *Foundations of Statistical Language Processing*. The MIT Press.

J.R. Newton and M. Sur. 2004. Plasticity of cerebral cortex in development. In G. Adelman and B.H. Smith, editors, *Encyclopedia of Neuroscience*. Elsevier, New York, 3rd edition.

J. Rissanen. 1989. *Stochastic Complexity in Statistical Inquiry*. World Scientific Publ. Co., New Jersey.

Z. Solan, D. Horn, E. Ruppin, and S. Edelman. 2005. Unsupervised learning of natural languages. *Proceedings of the National Academy of Sciences*, 102(33):11639–11634.

M. van Zaanen. 2000. Bootstrapping syntax and recursion using alignment-based learning. In P. Langley, editor, *Proc. ICML 2000*, pages 1063–1070. Stanford University, Morgan Kaufmann Publishers.