

Computational modeling of contexts and constructions

Extended abstract

Our research aims at developing computational models of the cognitive and social processes that underly language learning and use. At the cognitive level, we are interested in identifying the basic information processing principles that are, for instance, capable of producing gradually more abstract representations grounded in the experience of an individual. At the social level, we are interested in the processes of cultural evolution that lead into convergence and divergence in language use. We also wish to explore through computational modeling the connection between the cognitive and social level of language. In the following, we present four examples of our research activities:

- Unsupervised learning of morphology in which the use of machine learning is illustrated with limited use of context
- Learning constructions from text data
- Statistical modeling of contextual use of words using the self-organizing map and independent component analysis
- Associating linguistic expressions with external visual contexts

We begin by discussing some general aspects behind computational modeling in linguistics.

General remarks on (computational socio-cognitive) linguistics

Computational linguistics is an area in which computers have been used for a relatively long time as a research tool. Linguistics can be considered to be particularly interesting from the point of view of scientific practice and scientific representation because language is a central means for representing and communicating scientific results. In linguistics, the representational levels are intertwined, as illustrated in Figure 1. The community of language users (basically all human beings and in some sense also an increasing number of computational artifacts that process natural languages such as machine translation tools) produce and create language. By producing, we refer to the generation of linguistic expressions according to existing “rules and principles” including syntactic rules and lexical items. Creation refers to the fact that these above mentioned rules and principles are occasionally reformulated in their details by introducing new words to a language or by promoting new constructions that are gradually taken into common use (Honkela 2010).

From the methodological point view, a distinction into two basic paradigms in computational linguistics can be made. The first paradigm relies on the explicit encoding of linguistic knowledge by linguists based on their intuitions. The second paradigm is based on creating theories and models of language using statistical methods. The latter paradigm relies on the availability of large corpora that can be used to test the validity of hypotheses or to train the models.

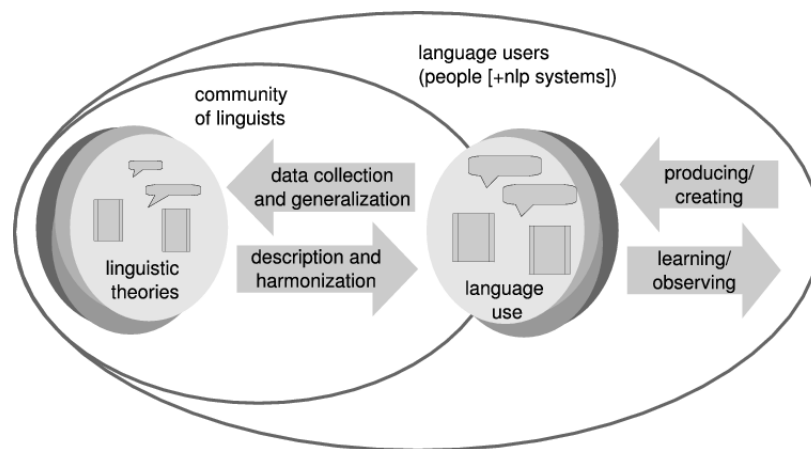


Figure 1. Social construction of language and linguistic knowledge (Honkela 2010).

The second paradigm can be further divided into two main alternatives. In the hypothesis-driven paradigm, predefined hypotheses are tested against the evidence provided by the corpus data. In the data-driven paradigm, statistical machine learning methods are used to build models of linguistic phenomena. The results of such model building depend on the representation of the linguistic input, methods and parameters used. In fact, both approaches can be viewed as hypothesis testing, the difference being in the number of hypotheses being tested and the nature of hypothesis generation. While in building scientific theories one is able to test only a small number of hypotheses, whereas a learning method may explore a space of thousands or even millions of hypotheses – the limit being the amount of data that is available. In the latter case the hypotheses are viewed as the *model space* where the learning takes place. Often the model space can be considered implicit as it is not easily understandable what the explored models are like.

In the following, we represent results based on the learning methods approach. This approach takes the focus to the (socio-)cognitive processes of language learning.

Unsupervised learning of morphology

Linguistic methods and automatic tools for retrieving morphological analyses for words, for example, based on the two-level morphology formalism (Koskeniemi 1983). However, these systems must be tailored separately for each language, which demands a large amount of manual work by experts. Moreover, specific tasks often require specialized vocabularies which must keep pace with the rapidly evolving terminologies. If it is possible to discover a morphology automatically from unannotated text, language and task independence are easier to achieve. We have demonstrated that it is possible to come up with a model that captures regularities within the set of observed word forms by observing the language data alone (Creutz and Lagus 2007). If a human were to learn a language in an analogous way, this would correspond to being exposed to a stream of large amounts of language without observing or interacting with the world where this language is produced. This is clearly not a realistic assumption about language learning in humans. However, Saffran et al. (1996) have shown that adults are capable of discovering word units rapidly in a stream of a nonsense language without any connection to meaning. This suggests that humans do use distributional cues, such as transition probabilities between sounds, in language learning. And these kinds of statistical patterns in language data can be

successfully exploited by appropriately designed algorithms.

The model family that we call Morfessor consists of various components that can be combined in different configurations. The model optimization criteria are expressed using the Maximum a Posteriori framework (Creutz and Lagus 2007). Morfessor segments the input words into units called morphs. As part of the learning process, a lexicon of morphs is constructed where information about both the distributional nature (usage) and form of each morph is stored. Usage relates to the distributional nature of the occurrence of morphs in words. Form corresponds to the string of letters that comprise the morph. An implementation of the method is freely available, and it is currently in active use for example in Finnish speech recognition research (Hirsimäki, Pylkkönen and Kurimo 2009).

Learning constructions from texts

A reasonable linguistic approach is offered by constructionist approaches to language, where language is viewed as consisting of constructions. The form component of the construction is not limited to a certain level of language as in many 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 will describe our first attempts 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 called Minimum Description Length (MDL). We have applied the developed method, for instance, to a corpus of stories told by 1–7 year old Finnish children, in order to look at constructions utilized by children (Lagus, Kohonen and Virpioja 2009).

Analysis of words in their linguistic contexts

Honkela, Pulkki and Kohonen (1995) presented an experiment in which a selection of 150 words from the English translation of Grimm fairy tales were analyzed based on the context statistics using the self-organizing map algorithm. In the resulting map, in which 150 most frequent words of the tales were included, the verbs and nouns in general and several subcategories of them emerged. An important consideration is that in the experiments the input did not contain any predetermined classifications. The results indicate that the text input as such, with the statistical properties of the contextual relations, is sufficient for automatic creation of meaningful implicit categories.

Lagus, Airola and Creutz (2002) analyzed the use of Finnish verbs to uncover possible conceptual spaces, and to study semantic similarities of verbs in actual language use. They examined the kinds of semantic or conceptual ordering qualities that appear to affect the distribution of features in the immediate context of a verb. With the unsupervised analysis based on the self-organizing map, emergent categories were found, such as manipulative actions in human relationships, start of action (with focus on will or intention), communication (especially with positive emotional information), and aggressive or destructive use of power. This result further strengthens the argumentation presented above as no predefined categories were in use, i.e., only statistical co(n)textual information was present in the input.

We have also explored the use of independent component analysis (ICA) for the automatic extraction of linguistic roles or features of words (Honkela, Hyvärinen and Väyrynen 2010). The extraction is based on the unsupervised analysis of text corpora. We contrast ICA with singular value decomposition (SVD), widely used in statistical text analysis, in general, and specifically in latent semantic analysis (LSA). The representations found using the SVD analysis cannot easily be interpreted by humans and do not thus have any direct or explicit linguistic relevance. In contrast, independent component analysis applied on word context data gives distinct features which reflect linguistic categories.

Analysis of multimodal contexts

A challenging problem is that how to computationally model the interrelated processes of understanding natural language, perceiving multimodal real world contexts and producing movement in them. Namely, in most cases computer systems processing symbols or language do not have access to the phenomena being referred to. In contrast, human beings can readily associate expressions with their non-linguistic experiences. As a direct consequence, the computational systems can only reason about the symbols themselves rather than about the meaning or external references of those symbols. This problem can be addressed in various ways. A traditional solution is to formalize the domain, such that the symbols used are defined in relation to each other to allow algorithmic analysis. However, if we want to process language as it is actually used by humans, we have to take into account that language is fully understood only through its use in linguistic and multimodal contexts. Another solution is to model the use of the symbols statistically.

In general, multimodal contexts are more naturally represented as patterns and signals which differ considerably from the discrete representation of symbols and expressions of symbolic languages. This finding has some important consequences on the methodology that is needed in the computational modeling of these phenomena and processes. At the experimental level, we have some preliminary results (such as Sjöberg et al. 2006), but, in general, this an area of ongoing work with long term objectives due to its complexity. The basic idea is that the computational processing of visual input finally takes place at the level of the original images and videos without the help of any manually generated labels. For the moment, there are only some limited and partial solutions for the automatic mapping between the visual and linguistic levels. Image analysis can also be supported by analyzing the direction of visual attention by analyzing gaze patterns (see, e.g., Puolamäki, Ajanki and Kaski 2008).

References

- Mathias Creutz, and Krista Lagus. Unsupervised Models for Morpheme Segmentation and Morphology Learning. *ACM Transactions on Speech and Language Processing*, 4(1): Article 3, January 2007.
- Teemu Hirsimäki, Janne Pylkkönen, and Mikko Kurimo. Importance of High-Order N-Gram Models in Morph-Based Speech Recognition. *IEEE Transactions on Audio, Speech & Language Processing* 17(4): 724-732, 2009.
- Timo Honkela. Directions for e-science and science 2.0 in human and social sciences. *Proceedings of MASHS 2010, Computational Methods for Modeling and Learning in Social and Human Sciences*, pages 119–134. Multiprint, 2010.
- Timo Honkela, Aapo Hyvärinen, and Jaakko Väyrynen. WordICA - Emergence of linguistic representations for words by independent component analysis. *Natural Language Engineering*, 16(3):277–308, 2010.

Timo Honkela, Ville Pulkki, and Teuvo Kohonen. Contextual relations of words in Grimm tales, analyzed by self-organizing map. *Proceedings of ICANN'95, International Conference on Artificial Neural Networks*, volume II, pages 3–7, Nanterre, France, 1995. EC2.

Kimmo Koskenniemi. Two-level Morphology. PhD thesis, University of Helsinki, 1983.

Krista Lagus, Anu Airola, and Mathias Creutz. Data analysis of conceptual similarities of Finnish verbs. In *Proceedings of the CogSci 2002, the 24th annual meeting of the Cognitive Science Society*, pages 566-571. Fairfax, Virginia, August 7-10, 2002

Krista Lagus, Oskar Kohonen, and Sami Virpioja. Towards unsupervised learning of constructions from text. *Proceedings of the Workshop on Extracting and Using Constructions in NLP of the 17th Nordic Conference on Computational Linguistics, NODALIDA*, May 2009. SICS Technical Report T2009:10.

Kai Puolamäki, Antti Ajanki, and Samuel Kaski. Learning to Learn Implicit Queries from Gaze Patterns. *Proceedings of International Conference on Machine Learning (ICML2008)*, pages 760-767, 2008.

Jenny R. Saffran, Richard N. Aslin, and Elissa L. Newport (1996). Statistical learning by 8-month-old infants. *Science*, 274, 1926–1928.

Mats Sjöberg, Ville Viitaniemi, Jorma Laaksonen, and Timo Honkela. Analysis of semantic information available in an image collection augmented with auxiliary data. *Proceedings of AIAI'06, Artificial Intelligence Applications and Innovations*, volume 204, pages 600–608. Springer, 2006.