Building a Finnish SOM-based ontology concept tagger and harvester

Seppo Nyrkkö
University of Helsinki
seppo.nyrkko@helsinki.fi

Abstract

I demonstrate here an experiment of word sense disambiguation method based on the Self-Organizing Map (SOM) and a pre-existing set of tools for analyzing text in Finnish. It is given a Semantic Web ontology as a reference model, and a related Finnish text corpus with sample term tagging related to the ontology concepts. The experiment is based on ”OntoR”, a previous experiment on SOM-based ontology term tagging for English. In this work the OntoR model is adapted to the Finnish language, and it is trained on a small text example with hand-picked concept annotations. This computational model can be considered useful for Information Retrieval and concept harvesting purposes in a specific domain where a limited training data set is available. The model adapted to Finnish text analysis stands on OMORFI and HFST morphological analysis, and uses the SOM-PAK library for unsupervised clustering, and ontology concept tagging and further for concept harvesting in Semantic Web ontology development.

Tiivistelmä

Kehitän luonnollisessa kielessä ilmenevien sanojen merkitysten erotteluun sopivaa automaattista koneoppivaa työkalua. Laskennallinen malli perustuu itseoppivaan karttaan (SOM, Self-Organizing Map) ja annettuun suomenkieliseen semanttisen webin ontologiaan. Malli oppii tunnistamaan käsitteiden ilmenemistä mallitekstistä, johon on annotoitu (tagattu) malliksi aiemmin laaditun ongologian käsitteitä. Koe liittyy aiemmin englanninkielisten käsitteiden taggaamiseen liittyvään OntoR-koejärjestelyyn joka tutki tekstisyötteessä ilmenevien termien liittämistä SOM-kartan soluihin malliksi annetun annotoidun tekstiesimerkin avulla. Tällainen malli oppii annetun käsitemallin huomattavan niukalla esimerkkiaineistolla ja sopii käyttökohteisiin joissa ei ole tarjolla riittävän suurta datamäärää syvän oppimisen neuroverkkomallin opettamiseksi. Suomenkielisen kokeen morfologisen analyysin pohjalla on OMORFI- ja HFST-työkalut. Koneoppimisen toteuttava SOM-kartta lasketaan SOM-PAK-ohjelmistopaketin avulla. Kehitettyä laskennallista mallia käytetään käsitteiden tunnistamisen lisäksi myös uusien ontologiakäsitteiden ehdokkaiden löytämiseksi.

1 Introduction

Plain word-based keywords might be misleading in some Information Retrieval purposes. The Semantic Web ontologies can provide enhanced results in information search when multiple taxonomies of terms and keywords are used in the document

Proceedings of the 4th International Workshop for Computational Linguistics for Uralic Languages (IWCLUL 2018), pages 18–25,
Helsinki, Finland, January 8–9, 2018. ©2018 Association for Computational Linguistics

database, for instance in medical or biological domain [1]. With automated ontology-concept tagging, a text database can be indexed incrementally to enhance queries defined by word-based examples or taxonomic identifiers. By referring to taxonomic concept identifiers in ontologies, both the tagging (indexing of terms and concepts) and Information Retrieval (search by terms or example phrases) can produce better precision in search results, compared to plain word-based keywords.

Text in Finnish is a challenge in Information Retrieval and automatic concept analysis due to its rich morphology and its marginal status in the existing forest of Semantic Web ontologies. By developing accessible and constitutive tools for analyzing morphologically rich languages, such as Finnish, the diverse work on automated and semi-automated concept tagging and multilingual ontology development will also become accessible. Also this way the methods developed for single languages can be evaluated in a foreign language or multilingual domain of the Semantic Web.

By using an automated concept tagging model, as aimed in the OntoR tool, it is possible to detect semantically significant features on tokens which link their usages to an ontology-based term. The detection of semantic features in the OntoR setup are based on a learning model, which is trained with data produced by a dependency parser program. The model described here also aims to disambiguate common words in special contexts where they are used as terms, as described in a Semantic Web ontology.

The utilized pre-processing software work with different levels (tokenization, lemmatisation, POS tagging and dependency parsing). The former English OntoR setup utilized the Stanford Parser PCFG model for English, but in this project I am using the OMORFI and HFST tools and UDPIPE tool adapted to the R environment.

Here the Finnish language is a very interesting challenge for dependency parsing since the word form disambiguation (e.g. lasta/lapsi) will be made in the R statistical programming environment, after the possible lemmas are parsed with HFST, but before estimation of the dependency graph with UDPIPE which will benefit from the lemma disambiguation. This project for adapting Finnish as the source language aims to normalize the set of pre-processing tools in a uniform model for analyzing concepts in multiple languages.

2 The Finnish OntoR experiment

This small demonstration aims to show how the Self-Organizing map method can work for unsupervised ontology term tagging and learning. The SOM is powerful in processing natural language since it can handle and learn on training data with a small set of significant outliers, and is robust in sense of accepting a noise component [2].

Neural network (NN) models for text-based learning are data-hungry when the model is trained with an unsupervised method. Word sense disambiguation require high-quality example training data, especially if the training data contains of homonyms and synonyms, reflecting real-life language. The concept detection method developed in the OntoR tool aims to be robust in cases of misspelled words and semantically equivalent alternatives by both a fuzzy character-based guessing (edit distance) technique and ontology-based semantic equivalence estimation. The out-of-lexicon words can be identified by a given synonym dictionary or applying typographic rules. Also the found words can be merged in the same concept by providing a synonym or a higher-class term (hypernym) in a Semantic Web ontology.

In contrast to most purely unsupervised neural network models, the OntoR model

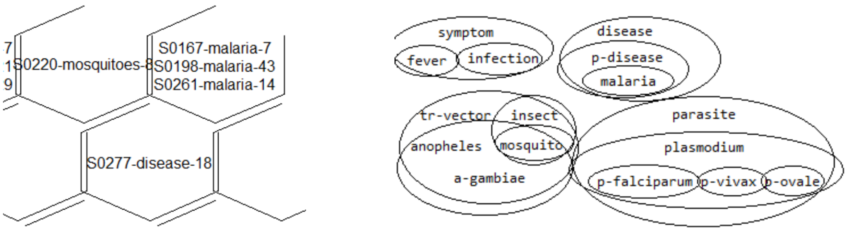


Figure 1: A detail of a Self-Organizing Map (SOM) and a related Venn diagram as a sample ontology for tagging terms

can be trained with a minimal training data set. For instance, the OntoR example development training data set for English contains a long Wikipedia article (9 500 words) and 100 related medical paper abstracts (24 000 words). This hand-picked development data set yielded a model capable of clustering a domain-related concept model, as sketched in Figure 1.

A previous, similar approach in English ontology concept term tagging has been done with the OntoR ontology term annotation tool, earlier developed in the EU MOLTO machine translation research project. Due to its open-to-develop nature it is very practical to extend its use by utilizing the existing Finnish morphological and syntactic analysis tools.

The OntoR was developed to use Stanford Parser for tokenization, lemmatization and extracting dependency information on natural language source text. The OntoR tool runs in the R statistical programming environment[3], using the CRAN library `som`, based on SOM-PAK[4], the Self-Organizing Map Program Package version 3.1.

2.1 Adapting OntoR to Finnish syntax

With this renovated experiment setup, I describe the required and planned steps to adapt the previously developed OntoR concept tagging model into a Finnish model for ontology concept tagging. As an extension to the previous OntoR experiment, I am now using HFST [5] and OMORFI [6] tools for Finnish corpus text lemmatization.

The developed syntactic analysis will use the Universal Dependencies (UD) data for Finnish [7]. For dependency arc computation, the Finnish OntoR model will be using the `udpipe` package for the R platform, instead of running the Stanford Parser model as an external process.

At the bootstrapping phase of adapting Finnish into the analysis model, I use 3-grams, which consists of the lemmatized base forms of the text node word, its previous and the following word. Practically this is done by adding ”left” and ”right” dependency arcs in the input sentence data. In a later step I intend to adapt the udpipe dependencies analysis developed in the UD project. This is expected to be equally powerful in expression, compared to the Stanford Parser PENN collapsed dependencies for English, which was used in the English OntoR setup.

In the development phase, a sample development corpus of 80 sentences were extracted from the Finnish wikipedia articles for Malaria and Protozoa (*fi:Alkueläimet*). A sample of this text in the OntoR environment can be seen in Figure 3. This aims at utilizing the same development ontology used for developing the English semantic model, shown in Figure 1.

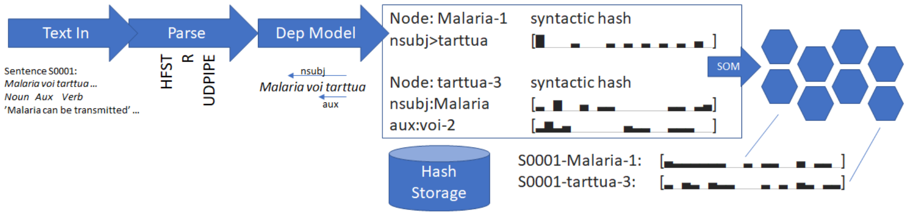


Figure 2: The workflow for producing semantic descriptors for syntactically analyzed text nodes.

2.2 Composing semantic feature vectors from syntactic arcs

Each *text node*, which is a specific occurrence of word in the source, gets a computed *semantic feature descriptor*. The which represents its observed syntactic neighborhood. The model generates *syntactic hash vectors* from syntactic dependencies, produced by the applied dependency parser and a random index generator. These hash vectors are composed into a weighted, distributional vector, used as the feature descriptor for the text node, which is practically the specific token in the sentence.

Since this is a probabilistic model, I chose random indexing as embeddings to represent lemmas and their typed syntactic dependencies. A random vector projection to a small dimension (20 at the first experiment) is applied to make computation affordable. The feature descriptors for text nodes are averaged from the set of their related syntactic hash vectors, and weighted by their inverse frequency. Similarly to the TF-IDF principle, a token occurring only a few times the weighting gives a large coefficient and commonly occurring tokens will get a smaller weight in the combined representation. The vector components are positive and L1-normed to sum of 1. A pipeline describing the feature generation process is shown in Figure 2.

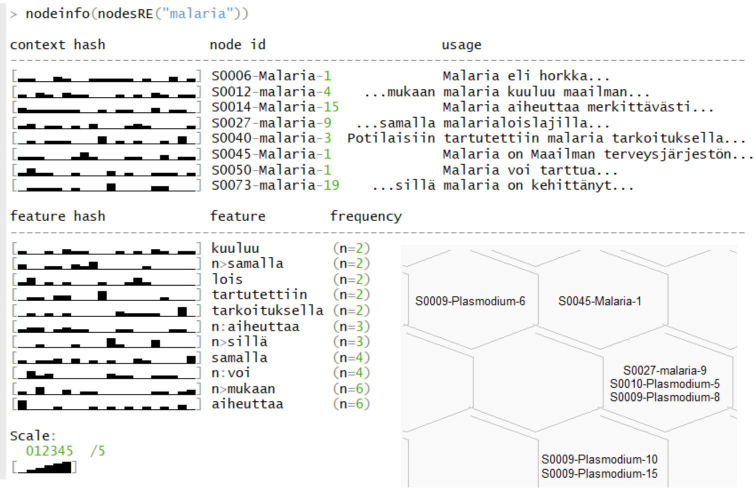
This distribution-based numeric representation has been chosen over Euclidean vector-space models (such as word2vec) due to the requirements of statistical analysis: The components must be able to be interpolated, summed and weighted, so that presence of any components may be measured in a combined feature descriptor. Also the difference between two semantic feature descriptors can be measured by a L1-distance or an entropy based distance such as IRad (information radius).

A tuple consisting of a dependency attribute and its head/dependent word builds an individual indexed syntactic hash. Also, the reverse arc and their endpoint words produce indexed syntactic arcs. This way, both the head and dependent ends of arcs are given unique features.

In Fig. 3 is shown a screen capture of the OntoR user interface, used for examining a computed model of an ontology-related text and a selected set of text nodes. The user interface produces a coarse bar chart of evaluated semantic feature descriptors for the text nodes and the related syntactic hashes used in the computation.

2.3 Self-Organizing Map representing an ontology

The SOM model proves to be powerful in unsupervised learning of multidimensional vector input and can handle input vector spaces with multiple dense clusters and sparse outlier data points. It adapts its clustering structure to wide-scale multidimensional variance in the data set, and is robust in terms of accepting a noise component



```
> nodeinfo(nodesRE("malaria"))

context hash               node id                        usage
------------------------------------------------------------------------------------------
[                        ] S0006-Malaria-1                Malaria eli horkka...
[                        ] S0012-malaria-4       ...mukaan malaria kuuluu maailman...
[                        ] S0014-Malaria-15               Malaria aiheuttaa merkittävästi...
[                        ] S0027-malaria-9      ...samalla malarialoislajilla...
[                        ] S0040-malaria-3  Potilaisiin tartutettiin malaria tarkoituksella...
[                        ] S0045-Malaria-1                Malaria on Maailman terveysjärjestön...
[                        ] S0050-Malaria-1                Malaria voi tarttua...
[                        ] S0073-malaria-19       ...sillä malaria on kehittänyt...

feature hash               feature         frequency
------------------------------------------------------------------------------------------
[                        ] kuuluu          (n=2)
[                        ] n>samalla       (n=2)
[                        ] lois            (n=2)
[                        ] tartutettiin    (n=2)
[                        ] tarkoituksella  (n=2)
[                        ] n:aiheuttaa     (n=3)
[                        ] n>sillä         (n=3)
[                        ] samalla         (n=4)
[                        ] n:voi           (n=4)
[                        ] n>mukaan        (n=6)
[                        ] aiheuttaa       (n=6)

Scale:
  012345   /5
[      ]
```

Figure 3: Screen shot from the OntoR environment, running in the R statistical programming environment. At the top, the semantic feature descriptors associated with token *Malaria* are printed. They represent the matching text nodes in the Finnish OntoR training data set, with a short word context of their usage. The node names are prefixed with a Snnnn- identifier which identifies the sentence number in the training data bank. The node names have a suffix -n which indicates the position of the token in the sentence, which is essential in cases of multiple word occurrences. Below are the syntactic hash vectors, which are used to build the semantic descriptors for text node related context. These are used in the composition of context descriptors, by weighted summing. An inverse frequency weighting is used so that an associated syntactic feature with low frequency (n) will cause a greater effect in the resulting context descriptor. Features occurring only once are not evaluated since they are taken only to provide noise to the training data. At this phase of development, a baseline lemmatisation is used instead of a dependency parse. *On the bottom right corner*: Some of the text nodes, aligned in the SOM space, as a result of the training process.

as part of the input. [2]

The OntoR setup demonstrates how ontology-based term structure is reflected on top the trained SOM map containing the keywords. A modified plot of the SOM map has been developed to explore the mapping of ontology term classes and super-classes over the machine-learned term model trained with the sample corpus. The SOM map can also be seen to reflect a Venn diagram representing an ontology concept space [8].

2.4 Observations on training the SOM classifier

The model can be given a sample ontology describing the domain of the training data set. The training ontology concepts are equipped with references to the training corpus. After the model is trained, the SOM model will reflect the found matching ontology concepts when a syntactic feature vector is presented to its feature space. Also, if a new term, a new spelling or synonym for an existing concept is detected, it is expected to appear near an existing concept tag in the SOM grid.

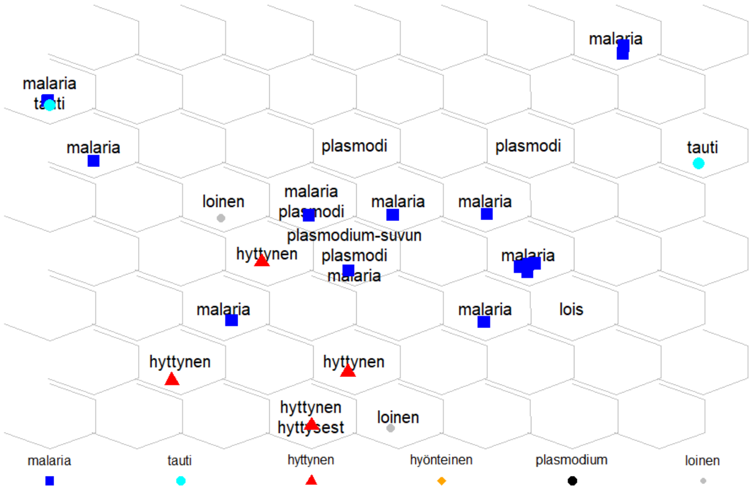


Figure 4: The SOM grid shown here demonstrates word contexts learnt from the development training data. The hexagonal cells are labeled with the associated terms, which are the words having a high frequency $n > 3$ in the training set. Each cell represents an averaged syntactic context, where the printed token is present. A cell may contain multiple tokens which appear in a similar syntactic context, and can be assumed to share some semantic features in common. Likewise, a token can appear in multiple cells, showing the syntactic context diversity of the specific token. An evaluation set of 20 sentences was separated from the development set of 80 sentences. The tokens selected for the evaluation are printed with their marking colors in the legend line at the bottom of the figure. The terms used in this evaluation plot were: *malaria*, *tauti* (en: disease), *hyttynen* (en: mosquito), *hyönteinen* (en: insect), *plasmodium*, *loinen* (en: parasite). These terms, evaluated in their sentence contexts are projected on a trained SOM model. The plot shows that the evaluation data points are located close to the ”target” clusters.

For development purpposes, I split the early development data into an evaluation set (20 sentences) and a development set (60 sentences) at random. A screen capture of a SOM-based term clustering at the development phase is shown in Fig. 4. In this example, three pairs of sub-terms and super-terms from the evaluation data set are plotted on the trained SOM model. The word form similarity measurements are disabled in this experiment. This shows that the evaluation data point features are well estimated without prior knowledge of the labels in the training data set, only based on their semantic feature vectors. Surprisingly, some word sense disambiguation happens even without the trained UDPIPE model attached.

The sample data used in the current development corpus is insufficient for numeric evaluation. Currently, at the time of writing, I am integrating the full syntacic analysis with the UDPIPE into the Finnish OntoR system, and also I am adding a larger corpus extracted from medical domain articles. This work seems to lead into a promisingly interesting evaluation of word sense disambiguation with the SOM and into further research on harvesting terms and introducing them as new ontology concepts.

3 Related work and development discussion

The SOM maps can also be seen learning Boolean elementary reasoning with logical statements in a restricted artificial language, when the model is trained by appropriate domain-specific text. The related work by Letosa et al. [8] supports this approach for using SOM in clustering the tagged concepts in given input in a restricted language. The boundaries between dense clusters can be seen as analogies to branches in taxonomy trees.

The similarity model based on the semantic descriptor vectors is very promising for containing contextual information on a word occurrence. Similar research on similarity measures on hash vectors has been recently done, as in work by Wang et al. [9]

Important work on automated and semi-supervised ontology population and extension has been done in the CultureSampo [10] project. Their model is also based on word distribution models on analyzed text which makes comparison to this work relevant.

There is also previous work on concept mining for the Semantic Web with SOM, for example the work by Honkela et al. [11], where the emergent structure of an organized SOM reflects the structure of underlying information, used in the training process. Their research also shows that multiple layers of superclass layers can be seen as different-sized nested zones on the SOM grid. This is analogous to the approach used in the concept classification (and further semantic disambiguation) in the OntoR project.

The expressiveness of Semantic Web ontologies and their language independent concept schema exceed the information in plain monolingual keyword-based taxonomies. Semantic web ontologies can contain relation attributes outside the superclass-subclass-taxonomy, such as *belongs-to* or *caused-by* relations. Ontology concepts may also be annotated with human readable description and machine-readable annotated for logical reasoning applications (e.g. through a SPARQL based schema). This suggests a need for research towards bridging the Semantic Web over multiple languages.

As a future step, an evaluation scheme for successful ontology concept tagging must be considered when developing the OntoR model further towards the pre-founded Finnish ontology structures, such as in the FinnONTO [12] project, and towards cross-linguistic concept tagging. This will also benefit the work of building Semantic Web ontologies and extending previously built monolingual ontologies to cover new languages and usages in cross-lingual Information Retrieval.

Acknowledgments

The development of the OntoR tool was initially funded by the EU MOLTO project at University of Helsinki. The experiment has been inspired by previous work on the SOM, lead by professor Timo Honkela. Timo has given beneficial feedback and ideas on this research subject.

References

[1] Jouni Tuominen, Nina Laurenne, and Eero Hyvönen. Biological names and taxonomies on the semantic web–managing the change in scientific conception. *The Semanic Web: Research and Applications*, pages 255–269, 2011.

[2] Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3):586–600, 2000.

[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017.

[4] Teuvo Kohonen, Jussi Hynninen, Jari Kangas, and Jorma Laaksonen. Som pak: The self-organizing map program package. *Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science*, 1996.

[5] Krister Lindén, Miikka Silfverberg, and Tommi Pirinen. Hfst tools for morphology–an efficient open-source package for construction of morphological analyzers. In *International Workshop on Systems and Frameworks for Computational Morphology*, pages 28–47. Springer, 2009.

[6] Krister Lindén, Erik Axelson, Senka Drobac, Sam Hardwick, Miikka Silfverberg, and Tommi A Pirinen. Using hfst for creating computational linguistic applications. In *Computational Linguistics*, pages 3–25. Springer, 2013.

[7] Sampo Pyysalo, Jenna Kanerva, Anna Missilä, Veronika Laippala, and Filip Ginter. Universal dependencies for finnish. In *Proceedings of the 20th Nordic Conference of Computational Linguistics, NODALIDA 2015, May 11-13, 2015, Vilnius, Lithuania*, number 109, pages 163–172. Linköping University Electronic Press, 2015.

[8] Jorge Ramón Letosa and Timo Honkela. Elementary logical reasoning in the som output space. In *International Conference on Artificial Neural Networks*, pages 432–437. Springer, 2010.

[9] Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji. Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927*, 2014.

[10] Tomi Kauppinen, Heini Kuittinen, Jouni Tuominen, Katri Seppälä, and Eero Hyvönen. Extending an ontology by analyzing annotation co-occurrences in a semantic cultural heritage portal. In *Proceedings of the ASWC 2008 Workshop on Collective Intelligence (ASWC-CI 2008), 3rd Asian Semantic Web Conference (ASWC 2008), Bangkok, Thailand*, pages 8–11, 2009.

[11] Timo Honkela and Matti Pöllä. Concept mining with self-organizing maps for the semantic web. In *WSOM*, pages 98–106. Springer, 2009.

[12] Eero Hyvönen, Kim Viljanen, Jouni Tuominen, and Katri Seppälä. Building a national semantic web ontology and ontology service infrastructure–the finnonto approach. In *European Semantic Web Conference*, pages 95–109. Springer, 2008.