An approach to unsupervised ontology term tagging of dependency-parsed text using a Self-Organizing Map (SOM)

Seppo Nyrkkö

Department of Digital Humanities, University of Helsinki
firstname.lastname@helsinki.fi

Abstract. I describe here a machine-learning estimation method for term tagging which can learn semantic disambiguation. The model is trained with a Semantic Web ontology, and a set of sample text documents with a set of concepts tagged, referring to the given ontology. The machine-learning method is based on creating numeric representations, or embeddings, which are based on dependency analysis of the syntactic environment of the word being analyzed. In contrast to many modern neural data-driven models, this model uses a less data-hungry unsupervised clustering method, the Self-Organizing Map (SOM). Based on the observations found with the experimental model, I suggest this can be utilized for populating ontologies with new concepts and terms, and for guessing the best matching ontology concepts for the found terms.

1 Introduction

Large amounts of written information flow in news, article databases and knowledge forums, and searching for required information often requires using proper keywords. Semantic Web ontologies describe a vocabulary of concepts and terms which are useful for Information Retrieval in their specified domain.

Ontologies can provide enhanced results in information search when multiple taxonomies of terms and keywords are used in composing a large document database. Such databases may cover for instance a multilingual, cultural or biological domain [1] where problems may be caused by diverse term variants, historical synonyms, misspellings and foreign terms.

Using automated content analysis based on machine learning, the amount of manual work in concept annotation and keyword tagging can be reduced. Automatic concept tagging makes it possible to apply ontology-based retrieval methods that combine keyword search with concept-based search [2]. This leads to a better coverage and quality compared to standard information retrieval.

I suggest here a method where a machine learning model is trained for semantic tagging. For demonstration purposes, a model is trained with a small annotated text, containing a set of examples of the terms described in the annotation ontology. In figure 1 a sample ontology used for the experiment is shown as a Venn diagram where separate and nested concept clusters are shown as graphical regions.

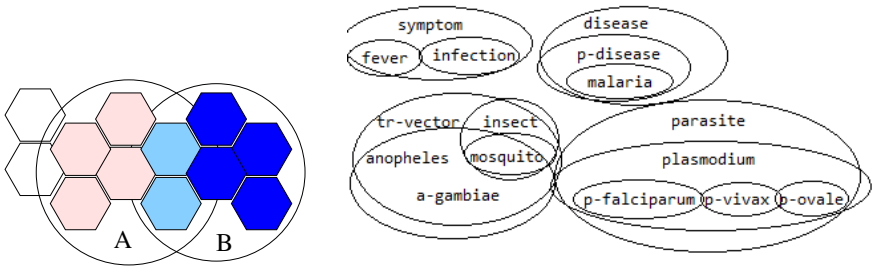

Fig. 1. *Left:* A visualization of a hexagonal Self-Organizing Map lattice (SOM) with data clusters of hypothetical concepts A and B. The SOM lattice has a tendency for clustering data points with similar features in cells next to each other. *Right:* A sample ontology, used for tagging terms in the experiment, shown as a Venn diagram.

2 The Method

The method intends to assist the process of adding semantic tags to individual sentences and paragraphs in new documents to be added to the database. The input text in new document is analyzed a sentence at a time, with a dependency parser. The semantic similarities between terms in new input and the reference text (training data) are estimated by the similarities in their syntactic dependencies in the new document.

A high syntactic similarity is considered as a possible semantic match. Furthermore, the method can be extended to detect a new term, without proper match. The approach also finds the closest match for an out-of-vocabulary term, that is not yet introduced in the current ontology.

By using an unsupervised machine learning method such as the Self-Organizing Map (SOM) we can even give a comprehensive, visual impression of the collection of articles available in the text database. A SOM is a Neural Network model that is different from most modern neural network architectures. It is less data-hungry and it is tolerant to noise in the training data [3]. This way, it can also classify rare term occurrences that have no exact match in the training data set, by guessing the best partial match based on the syntactic features of the term. This makes it an interesting alternative model to learning features for terms, associated with a set of ontology concepts.

3 Sample experiment

In the experiment, the sentences of text corpora are processed with the Stanford Parser (Penn PCFG dependency model for English). The sentences are tokenized as part of the dependency parsing process. Each token (in its actual word form) in the sentences are indexed in the training sentence bank. The dependency arcs and bi-arcs are extracted from the parse output, and each arc forms a

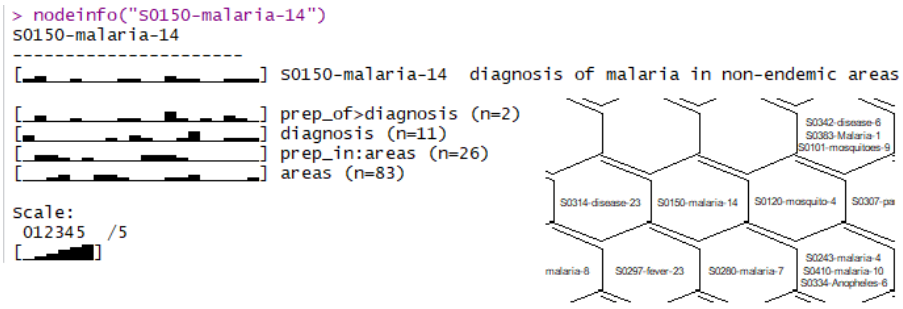

Fig. 2. The user interface of the OntoR tool, showing a cell of the Self-Organizing Map (SOM), a matching token as a data point and its syntactic descriptor components. The sample text was based in the Malaria text downloaded from Wikipedia (May 2017) and set of 200 pubmed article abstracts (100 of keyword *mosquito* and 100 of *malaria*).

feature descriptor on the tagged word. The arcs are bi-directional so that one dependency is tagged on the head and dependent word. The semantic features for individual word tokens are random projected indexes of the features produced by the Stanford Parser model. The syntactic context representation is very similar to the one in Dependency-Based Word Embeddings as in [4].

The experimental OntoR tool was developed in the R statistical programming environment, using the CRAN library `som`, based on *SOM-PAK*, the Self-Organizing Map Program Package (version 3.1) [5]. A screen shot from the OntoR user interface (2) demonstrates how ontology-based term structure is reflected as a 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 term model trained with the sample corpus.

The areas in the outcome SOM grid show the taxonomical hierarchy that can be seen in the mapping of ontology-terms in the unsupervised model representing the training corpus. Multiple clusters were seen with both the subterms and terms categorized in the same map cell and their neighborhood. This supports the earlier work hypothesis that a data point cluster with an internal topology, or a structure, has a strong tendency to distribute over multiple adjacent cells over the SOM lattice.

4 Related Work and Discussion

The WebSOM project[6] inspired work towards unsupervised term learning and classification with the use of Self Organizing Map, which works and learns on Internet sourced text articles, and extracts topics based on the tokens found in the text articles. Also the work by Tanev et al [7] describes the main paradigms on weakly supervised ontology population, one being the term pattern related method and the another being the context sensitive triggering. The approach

described here is a contextual extension to the WebSOM model since it adds syntactic dependencies as additional information over the tokens found in the text. In this work, the suggested method for mapping concepts occurring in text into the SOM grid will analogously support automatic tagging of new term candidates in document databases. This seems applicable especially for hyponyms (terms for subclasses) and synonyms for previously categorized terms. In the following phase of the experiment, the internal weighting parameters for building numeric embeddings from syntactic analysis will be evaluated and analyzed in contrast to using plain word-based embeddings.

This method can also be seen applicable in weakly supervised ontology concept population for adding new term candidates, since the presence of some rare terms occurrences were found in distinct areas of the SOM map in the experiment. This aim to use the SOM in concept mining is also supported by work by Honkela and Pöllä [8]. The set of ontologies used with OntoR is not restricted to a medical domain, as seen with the sample experiment. The used ontologies can even cover multiple topics, for instance, history, politics, science and culture.

Acknowledgments: Research and development of the method and the OntoR tool have been supported by the MOLTO EU project and Whitelake Software Point. The suggested model and inspection of the methods described here have been developed and supported with feedback from professor Timo Honkela and the Research Seminar in Language Technology held at University of Helsinki.

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. Minna Tamper, Petri Leskinen, Esko Ikkala, Arttu Oksanen, Eetu Mäkelä, Erkki Heino, Jouni Tuominen, Mikko Koho, and Eero Hyvönen. Aatos–a configurable tool for automatic annotation. In *International Conference on Language, Data and Knowledge*, pages 276–289. Springer, 2017.
3. Juha Vesanto and Esa Alhoniemi. Clustering of the self-organizing map. *IEEE Transactions on neural networks*, 11(3):586–600, 2000.
4. Omer Levy and Yoav Goldberg. Dependency-based word embeddings. In *ACL (2)*, pages 302–308, 2014.
5. 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.
6. T Honkela, S Kaski, T Kohonen, and K Lagus. Self-organizing maps of very large document collections: Justification for the websom method. In *Classification, Data Analysis, and Data Highways*, pages 245–252. Springer, 1998.
7. Hristo Tanev and Bernardo Magnini. Weakly supervised approaches for ontology population. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.
8. Timo Honkela and Matti Pöllä. Concept mining with self-organizing maps for the semantic web. In *WSOM*, pages 98–106. Springer, 2009.