Available online at www.sciencedirect.com

ScienceDirect

Computer Speech & Language 46 (2017) 113–135

www.elsevier.com/locate/csl

A time-sensitive historical thesaurus-based semantic tagger for deep semantic annotation☆

Scott Piao[a,*], Fraser Dallachy[b], Alistair Baron[a], Jane Demmen[a], Steve Wattam[a], Philip Durkin[c], James McCracken[c], Paul Rayson[a], Marc Alexander[b]

[a] *Lancaster University, Lancaster LA1 4WA, United Kingdom*
[b] *University of Glasgow, Glasgow G12 8QQ, United Kingdom*
[c] *Oxford University Press, Oxford OX2 6DP, United Kingdom*

Received 22 July 2016; received in revised form 5 February 2017; accepted 30 April 2017
Available online 17 May 2017

Abstract

Automatic extraction and analysis of meaning-related information from natural language data has been an important issue in a number of research areas, such as natural language processing (NLP), text mining, corpus linguistics, and data science. An important aspect of such information extraction and analysis is the semantic annotation of language data using a semantic tagger. In practice, various semantic annotation tools have been designed to carry out different levels of semantic annotation, such as topics of documents, semantic role labeling, named entities or events. Currently, the majority of existing semantic annotation tools identify and tag partial core semantic information in language data, but they tend to be applicable only for modern language corpora. While such semantic analyzers have proven useful for various purposes, a semantic annotation tool that is capable of annotating deep semantic senses of all lexical units, or all-words tagging, is still desirable for a deep, comprehensive semantic analysis of language data. With large-scale digitization efforts underway, delivering historical corpora with texts dating from the last 400 years, a particularly challenging aspect is the need to adapt the annotation in the face of significant word meaning change over time. In this paper, we report on the development of a new semantic tagger (the Historical Thesaurus Semantic Tagger), and discuss challenging issues we faced in this work. This new semantic tagger is built on existing NLP tools and incorporates a large-scale historical English thesaurus linked to the Oxford English Dictionary. Employing contextual disambiguation algorithms, this tool is capable of annotating lexical units with a historically-valid highly fine-grained semantic categorization scheme that contains about 225,000 semantic concepts and 4,033 thematic semantic categories. In terms of novelty, it is adapted for processing historical English data, with rich information about historical usage of words and a spelling variant normalizer for historical forms of English. Furthermore, it is able to make use of knowledge about the publication date of a text to adapt its output. In our evaluation, the system achieved encouraging accuracies ranging from 77.12% to 91.08% on individual test texts. Applying time-sensitive methods improved results by as much as 3.54% and by 1.72% on average.

Keywords: Semantic annotation; Natural language processing; Historical thesaurus; Semantic lexicon; Corpus annotation; Language technology

☆ This paper has been recommended for acceptance by Pascale fung.

* Corresponding author.
E-mail address: s.piao@lancaster.ac.uk (S. Piao), fraser.dallachy@glasgow.ac.uk (F. Dallachy), a.baron@lancaster.ac.uk (A. Baron), j.e.demmen1@lancaster.ac.uk (J. Demmen), s.wattam@lancaster.ac.uk (S. Wattam), philip.durkin@oup.com (P. Durkin), james.mccracken@oup.com (J. McCracken), p.rayson@lancaster.ac.uk (P. Rayson), marc.alexander@glasgow.ac.uk (M. Alexander).

http://dx.doi.org/10.1016/j.csl.2017.04.010

1. Introduction[1]

Semantic analysis of natural language data is a relevant task for a wide range of research areas and practical applications, such as natural language processing, text mining, corpus linguistics and data science. Numerous semantic annotation tools have been developed to carry out various levels of semantic analysis, such as document topics, named entities, temporal information, and so on. For example, some tools are designed to identify the topic or themes of given texts (Allan, 2012), and some are designed to extract specific partial information, such as types of named entities, categories of relations between the specific named entities, and/or types of events (Miwa et al., 2012; Rizzo and Troncy, 2012; Weston et al., 2013). Another group of semantic annotation tools are designed to identify semantic categories of all lexical units based on a given classification scheme, which can support a deep comprehensive semantic information analysis and extraction from language data. The latter task entails richer semantic lexical resources and a deeper level of sense disambiguation, and hence presents tough challenges. Our work presented in this paper addresses the issue of a semantically rich text analytical system.

Over recent years, various semantic lexical resources and semantic annotation tools have been developed, such as EuroWordNet (Vossen, 1998) and the UCREL (University Centre for Computer Corpus Research on Language) Semantic Analysis System (USAS) (Rayson et al., 2004), and they have played an important role in developing intelligent natural language processing (NLP) and Human language technology (HLT) systems. For example, the USAS semantic tagger has been applied in a variety of studies, including empirical language studies at the semantic level (Klebanov et al., 2008; Ooi et al., 2007; Potts and Baker, 2013; Rayson et al., 2004), studies in information technology (Doherty et al., 2006; Nakano et al., 2005; Volk et al., 2002), software engineering (Chitchyan et al., 2006; Taiani et al., 2008) and others (Balossi, 2014; Gacitua et al., 2008; Hancock et al., 2013; Markowitz and Hancock, 2014; Semino et al., 2015).

In this paper, we present our work in designing, developing and evaluating the accuracy of a new semantic tagger: the "Historical-Thesaurus-based Semantic Tagger" (henceforth HTST). The purpose of this tool is to annotate all lexical units of texts with a fine-grained semantic categorization scheme provided by a very large-scale and high-quality English historical thesaurus (Kay et al., 2016 [2009]) (detailed further in the next section).

2. Related work

In recent years, researchers have devoted a great deal of effort to the development of various semantic annotation tools of natural language data. In particular, various lexical knowledge bases have been used to assign semantic concepts and categories to words and other types of lexical units in text. For example, WordNet is widely used for such a purpose, as demonstrated by the collection of WordNet Sense annotated corpora at the website http://globalwordnet.org/wordnet-annotated-corpora (last accessed 19 April 2016). A similar approach has been used for developing a more semantic field oriented semantic tagger, USAS, at UCREL (Lancaster University, UK; http://ucrel.lancs.ac.uk/usas), which is based on semantic lexicons containing lexical units classified with a set of pre-defined coarse-grained semantic fields rather than grouped by fine-grained word senses as in WordNet.

A significant amount of effort has been dedicated in previous research to word sense disambiguation, in particular in the SensEval series of events (Evaluation Exercises for the Semantic Analysis of Text; http://www.senseval.org), and more recently this has widened out (in SemEval and *SEM) to encompass other elements of computational analysis of meaning. Although, in some cases these do use existing sense inventories (e.g. BabelNet), generally the sense inventory is induced or clustered from a training set. Corpus-based distributional semantic models and word embeddings are now proving a very popular approach but generally still conflate different meanings of words under a single vector representation. In some works (Iacobacci et al., 2015), this limitation is starting to be addressed, but so far no research has been able to leverage meaning change over time, and this is obviously key for semantically annotating

[1] Abbreviations: CLAWS=Constituent Likelihood Automatic Word-tagging System; EEBO=Early English Books Online; EModE=Early Modern English; GATE=General Architecture for Text Engineering; HTST=Historical Thesaurus Semantic Tagger; MWE=MultiWord Expression; NLP=Natural Language Processing; OE=Old English; OED=Oxford English Dictionary; POS=Part-of-Speech; SAMUELS=Semantic Annotation and Mark-up for Enhancing Lexical Searches; UCREL=University Centre for Computer Corpus Research on Language; USAS=UCREL Semantic Analysis System; VARD=Variant Detector Software.