Available online at www.sciencedirect.com

ScienceDirect

Lingua 157 (2015) 153–172

Lingua

www.elsevier.com/locate/lingua

Monitoring polysemy: Word space models as a tool for large-scale lexical semantic analysis

K. Heylen [*], T. Wielfaert, D. Speelman, D. Geeraerts

QLVL – KU Leuven, University of Leuven, Belgium

Received 4 February 2014; received in revised form 2 December 2014; accepted 3 December 2014
Available online 3 February 2015

Abstract

This paper demonstrates how token-level Word Space Models (a distributional semantic technique that was originally developed in statistical natural language processing) can be developed into a heuristic tool to support lexicological and lexicographical analyses of large amounts of corpus data. The paper provides a non-technical introduction to the statistical methods and illustrates with a case study analysis of the Dutch polysemous noun 'monitor' how token-level Word Space Models in combination with visualisation techniques allow human analysts to identify semantic patterns in an unstructured set of attestations. Additionally, we show how the interactive features of the visualisation make it possible to explore the effect of different contextual factors on the distributional model.

Keywords: Distributional models; Lexical semantics; Statistical analysis; Visual analytics

1. Introduction

Compared to other linguistic disciplines, corpus-based analyses have a strong and long tradition in lexical semantics. Ever since the rise of philology and the emergence of large-scale dictionary projects in the 19th century, lexical semanticians have relied on contextual clues in attested language use to infer and organise the different senses and uses of a word. And while in the 1950s syntax research turned away from usage data, the ideas of John Rupert Firth (1957), Zelig Harris (1954) and Warren Weaver (1955) led to approaches that saw real language data as the natural empirical basis for semantic descriptions. (For a more extended history of recent corpus-linguistic approaches to lexical semantics, see Geeraerts, 2010: 165–178.) Initially, collecting and analysing corpus data was mainly manual labour, but with the advent of computers and ever larger electronic corpora, lexicologists and lexicographers now have enormous amounts of naturally occurring usage data available to base their descriptive work on. To analyse this wealth of data, scholars of lexical semantics widely use statistical analysis tools. More specifically, statistical methods have been introduced to facilitate two distinct steps in the analysis. On the one hand, statistical methods are used for identifying contextual clues in the corpus data that are indicative of a given lexeme's meaning. These include co-occurring words (collocations) and syntactic patterns (colligations). On the other hand, statistical approaches are employed for classifying the occurrences of a lexeme into distinct usages and senses based on these contextual clues.

[*] Corresponding author. Tel.: +32 16329998; fax: +32 16324713.
E-mail address: kris.heylen@kuleuven.be (K. Heylen).

http://dx.doi.org/10.1016/j.lingua.2014.12.001
0024-3841/© 2014 Elsevier B.V. All rights reserved.

Table 1
Methods of sense identification in lexical semantics.

	Identifying contextual clues	Classifying occurrences
Philology	Manual	Manual
Collocation analysis	Statistical	Manual
Behavioural profiles	Manual	Statistical
Word Space Models	Statistical	Statistical

The first approach (i.e. the introduction of statistical methods for the identification of contextual clues) has been mainly associated with the British tradition in corpus linguistics. Pioneered by John Sinclair (1991), this approach described lexical meaning as a function of the typical words (collocations) and syntactic patterns (colligations) that a word co-occurs with. Church and Hanks (1989) introduced statistical measures (*t*-score), to identify salient and informative collocations and colligations based on frequency distributions in text. These measures were subsequently refined (see Evert, 2004 and Wiechmann, 2008 for an overview), and they are now widely used in various linguistic subdisciplines.

The second approach (i.e. statistical clustering of usages) has an outspoken presence in recent developments in Cognitive Semantics. Specifically, the so-called Behavioural Profile approach has introduced multivariate statistical techniques to classify occurrences of a word automatically into distinctive senses and usages, based on corpus evidence.[1] Gries (2006) uses hierarchical cluster analysis to group occurrences of the verb *to run* into different senses based on contextual features like transitive use or co-occurring spatial prepositions. Glynn (2010) applies Correspondence Analysis to visualise how occurrences of the verb *to bother* are grouped into distinct usages based on syntactic behaviour and semantic characteristics like affect.

Interestingly, these statistical methods have been used independently of each other in these different traditions. Collocation-based analyses have statistically automated the identification of contextual clues but leave the classification of occurrences and typical contexts to manual analysis. Almost as a mirror image, behavioural profile analyses have statistically automated the classification of a lexeme's occurrences and typical contexts into senses, but predominantly use datasets with manually coded contextual features as input. Table 1 classifies different approaches in lexical semantics by whether they identify contextual clues and senses manually or through statistical analysis. Whereas classical philological studies did both steps manually, collocation studies and behavioural profile analyses have each by and large automated one of the two steps but not both.

In this paper, we will introduce Word Space Models (a.k.a. Semantic Vector Space Models) as a logical extension of the statistical state-of-art in support of lexical semantic analysis: a technique that essentially combines collocational measures and multivariate methods in a systematic way to explore lexical semantic structure in large corpora, it completes the pattern that emerges from the introduction of statistical tools in corpus-based lexical semantics, as summarised in Table 1.

The introduction of statistical methods to both steps in the data analysis process is, we think, not only a logical extension, but also a necessary one and this, for two reasons. First, additional support of statistical pattern finding techniques is the only way for lexicologists and lexicographers to cope with the data deluge that they face as they pursue their traditional descriptive work. Given the available wealth of corpus data, it is simply unfeasible to hand-code, classify or describe thousands upon thousands of concordances of a lexeme. Statistical data analysis can help to take a representative sample of different usages for further scrutiny. Secondly, so-called Big Data also suggests an extension of the traditional focus of lexicological and lexicographical work: the Big Data environment allows scholars to investigate trends and patterns that could not be studied in smaller corpora, e.g. the spreading of new words or new usages of existing words through social networks. Data mining techniques are indispensable to monitor this type of trends. As Word Space Models are already the principal technique for handling lexical semantics in Computational Linguistics, we suggest that they can also provide support for the lexicologist and the lexicographer in their traditional and new descriptive tasks: for the traditional task of describing individual words or lexical fields, support is needed for staying on top of the abundance of data, and for the new task of describing trends and developments, appropriate quantitative techniques need to be developed.

In this paper, we will focus on how a semantic space approach can support a lexicological analysis of polysemy in large corpora. It should be noted, though, that we are explicitly not presenting the Semantic Vector Space approach as a ready-made, stable technique. Rather, we will argue that in its current, computational linguistic implementation, the technique is

[1] Within the behavioural profile approach, there are also studies of lexical alternations, for which other multivariate methods like regression are available. However, since our focus is on the polysemy of 1 lexeme, rather than the alternation between multiple lexemes for the same concept, we do not go further into these.