Conceptual Exploration of Semantic Mirrors

Uta Priss and L. John Old

Napier University, School of Computing
{u.priss, j.old}@napier.ac.uk

Abstract. The "Semantic Mirrors Method" (Dyvik, 1998) is a means for automatic derivation of thesaurus entries from a word-aligned parallel corpus. The method is based on the construction of lattices of linguistic features. This paper models the Semantic Mirrors Method with Formal Concept Analysis. It is argued that the method becomes simpler to understand with the help of FCA. This paper then investigates to what extent the Semantic Mirrors Method is applicable if the linguistic resource is not a high quality parallel corpus but, instead, a medium quality bilingual dictionary. This is a relevant question because medium quality bilingual dictionaries are freely available whereas high quality parallel corpora are expensive and difficult to obtain. The analysis shows that by themselves, bilingual dictionaries are not as suitable for the Semantic Mirrors Method but that this can be improved by applying conceptual exploration. The combined method of conceptual exploration and Semantic Mirrors provides a useful toolkit specifically for smaller size bilingual resources, such as ontologies and classification systems. The last section of this paper suggests that such applications are of interest in the area of ontology engineering.

1 Introduction

Dyvik (1998, 2003, 2004) invented the "Semantic Mirrors Method" as a means for automatic derivation of thesaurus entries from a word-aligned parallel corpus. His online interface[1] uses a parallel corpus of Norwegian and English texts, from which users can interactively derive thesaurus entries in either language. A feature set is derived for each sense of each word. The senses then form a semi-lattice based on inclusion and overlap among feature sets. Priss & Old (2004) note (without providing any details) that Dyvik's method is similar to certain concept lattices derived from monolingual lexical databases. The Semantic Mirrors Method is briefly described in section 2 of this paper. Section 3 explains how the Semantic Mirrors Method can be represented with respect to Formal Concept Analysis (FCA). We believe that the Semantic Mirrors Method is of general interest to the FCA community because there may be other similar applications in this area.

In section 4, the FCA version of the Semantic Mirrors Method from section 3 is applied to an English-German dictionary. An advantage of using bilingual dictionaries instead of parallel corpora is that bilingual dictionaries are freely available on the Web whereas word-aligned parallel corpora are expensive. A disadvantage of using bilingual

[1] http://ling.uib.no/~helge/mirrwebguide.html

B. Ganter and R. Godin (Eds.): ICFCA 2005, LNAI 3403, pp. 21–32, 2005.

dictionaries is that the semantic information which can be extracted from them is less complete, at least with respect to the creation of Semantic Mirrors. Therefore, in section 5 of this paper we analyse how conceptual exploration (cf. Stumme (1996)) can be used to improve the incomplete information extracted from bilingual dictionaries. Even though conceptual exploration is a semi-automated process, we believe that in combination with the Semantic Mirrors Method, this approach has potential applications with respect to ontology merging as described in section 6.

This paper attempts to provide sufficient details of the Semantic Mirrors Method to be understandable for non-linguists, but it is assumed that readers are familiar with the basics of FCA, which can be found in Ganter & Wille (1999).

2 The Semantic Mirrors Method

The Semantic Mirrors Method intends to extract semantic information from bilingual corpora, which are large collections of texts existing in two languages and which are aligned according to their translations. The assumption is that if the same sentence is expressed in two different languages, then it should be possible to align words or phrases (or "lemmata") in one language with the corresponding words or phrases in the other language. This word alignment is not trivial because languages can differ significantly with respect to grammar and syntactic ordering. Computational linguists have developed a variety of statistical algorithms for such word-alignment tasks. These algorithms perform with different degrees of accuracy. One of Dyvik's interfaces allows for users to vary the parameters used in these algorithms to explore their impact on the extracted Semantic Mirrors. For comparison, Dyvik has also experimented with manually aligned corpora[2]. For the purposes of this paper, only the resulting lists of aligned translations are of interest. The quality or accuracy of the word alignment algorithms are not discussed in this paper.

2.1 Step 1

Once a bilingual corpus is word-aligned, one can select a word in either language and list all translations of that word occurring in the corpus. These lists of words and their respective lists of translations form the basis of the Semantic Mirrors Method. Dyvik (2003) calls the set of translations of a word a from language A its "(first) t-image" in language B. One can then form the t-images (in language A) of the t-image (in language B) of word a from language A. This set of sets is called the "inverse t-image of a". This algorithm of collecting the translations of the translations of a word has been mentioned by other authors (for example, Wunderlich (1980)) and is called the "plus operator" by Priss & Old (2004). This algorithm presents the first step of Dyvik's Mirrors Method. In contrast to this first step which has independently been discovered by different authors, to our knowledge, the next steps of the Semantic Mirrors Method are unique to this method.

[2] `http://ling.uib.no/~helge/mirrwebguide.html\#bases`