Building a Chinese-English Mapping Between Verb Concepts for Multilingual Applications

Bonnie J. Dorr[1], Gina-Anne Levow[1] and Dekang Lin[2]

[1] Institute for Advanced Computer Studies
University of Maryland
College Park, MD, 20742
{bonnie,gina}@umiacs.umd.edu
[2] Department of Computing Science
University of Alberta
Edmonton, Alberta, Canada, T6G 2H1
lindek@cs.ualberta.ca

Abstract. This paper addresses the problem of building conceptual resources for multilingual applications. We describe new techniques for large-scale construction of a Chinese-English lexicon for verbs, using thematic-role information to create links between Chinese and English conceptual information. We then present an approach to compensating for gaps in the existing resources. The resulting lexicon is used for multilingual applications such as machine translation and cross-language information retrieval.

1 Introduction

With the advent of the web and increasingly more global interconnectivity, the need for online multilingual information has grown significantly in the last 5–10 years. This is accompanied by a growing need for rapid construction of lexical resources. Creating resources by human labor alone has become infeasible, thus motivating the development of automatic and semi-automatic approaches to resource acquisition. This paper addresses large-scale construction of a Chinese-English lexicon for verbs, including an approach to compensating for gaps in the existing resources.

The lexicons resulting from our acquisition approach are used for semantic analysis in applications such as machine translation and cross-language information retrieval. The importance of semantic analysis in either of these two applications is clear when one considers the degree of inaccuracy that might result from using a weak alternative, such as access to a bilingual word list.

Our starting point is an existing classification of English verbs called EVCA (English Verbs Classes and Alternations) [?]. We couple this with a Chinese conceptual database called HowNet [?,?,?] (http://www.how-net.com), from which we extract thematic-role information (e.g., a mapping between the HowNet "Patient" and the EVCA-based "Th(eme)") to create links between Chinese and

English conceptual information. HowNet currently contains no English translations; thus, we also use a large machine-readable Chinese-English dictionary called Optilex to produce candidate English translations.[1] Although later versions of HowNet are expected to include the English translations, these are not openly available—only the binary versions have been promised and these will be accessible solely through the use of (purchasable) HowNet software. Moreover, we expect our techniques to be generally applicable to *other* foreign language semantic hierarchies where English translations are not available. We predict this will occur more and more frequently, as online (non-bilingual) linguistic resources continue to be made available in multiple languages.

Several researchers have investigated the problem of assigning class-based senses to verbs [?,?,?] using a variety of online resources including Longman's Dictionary of Contemporary English (LDOCE) [?], EVCA [?], and WordNet [?]. Translation of English classes into other languages has proven difficult [?,?,?], but regularities between different language classifications can be found in some online resources [?,?,?].

This work extends previous work which used a concept space to produce a hierarchical organization of Chinese verbs [?]. We adopt a technique that is similar in flavor to that of [?] for partitioning English verbs into refined classes using WordNet, with the following extensions: (1) The use of the entire EVCA database rather than a small set of verbs (the *break* class); (2) The provision of a thematic-role based filter for a more refined version of verb-class assignments; (3) Concept alignment across two different language hierarchies (Chinese and English); and (4) Mappings between Chinese and English thematic roles.

This work relies on an augmented set of EVCA classes which include 26 new classes [?]. There are 500 total classes in the extended set, each hand-tagged with semantic representations, thematic-role information, and WordNet synset numbers. We will demonstrate that it is possible to produce a lexicon by associating 709 Chinese HowNet concepts with 500 EVCA classes, with a clear concept-to-class correspondence in a large majority of the cases.[2]

Figure 1 illustrates the relation between existing resources and the mappings we produced. Solid lines represent pre-existing mappings; dotted lines are ones resulting from the application of our techniques. The most critical of these is the one labeled θ-roles (shorthand for "thematic roles"), which associates EVCA classes with HowNet Concepts. The remaining two dotted-line mappings are "transitive closure" biproducts of the other mappings: Once the thematic-role mapping associates EVCA verbs with HowNet verbs, each HowNet verb is associated with Optilex-based English *glosses* (translations) and WordNet 1.6 Senses.

---

[1] Optilex is a large (600k entries) machine-readable version of the CETA Chinese-English dictionary, licensed from the MRM corporation, Kensington, MD.

[2] HowNet contains 815 verb HowNet concepts altogether. However, we are not including the 106 HowNet concepts that are not associated with any Chinese words; these are "higher level" conceptual nodes with no Chinese realization (e.g., V.1 |static|).

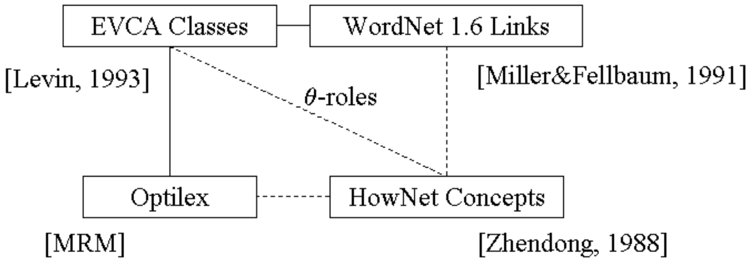

Fig. 1. Relation Between Existing Resources and New Mappings

We will describe how these correspondences are derived and we will show how this process has provided a framework for compensating for gaps in our online resources.

2 Multilingual Applications

The semantic representations produced semi-automatically for our multilingual resources are used in machine translation (MT) and cross-language information retrieval (CLIR) applications. Both applications rely on the use of a parser for mapping the input sentence into a syntactic tree. The parser output is semantically analyzed, producing an encoding of semantic and argument-structure information.

The MT approach is interlingual, where the target-language lexicon is searched for appropriate lexical items matching argument-structure information [?]. A screen snapshot of a MT example is shown in Figure 2. The CLIR approach relies on the same interlingual representation to transform a user's query into the document language for information retrieval [?,?].

[3] The Chinese verbs are additionally associated (for free) with WordNet senses from our previously tagged EVCA verbs. More details are given in [?].

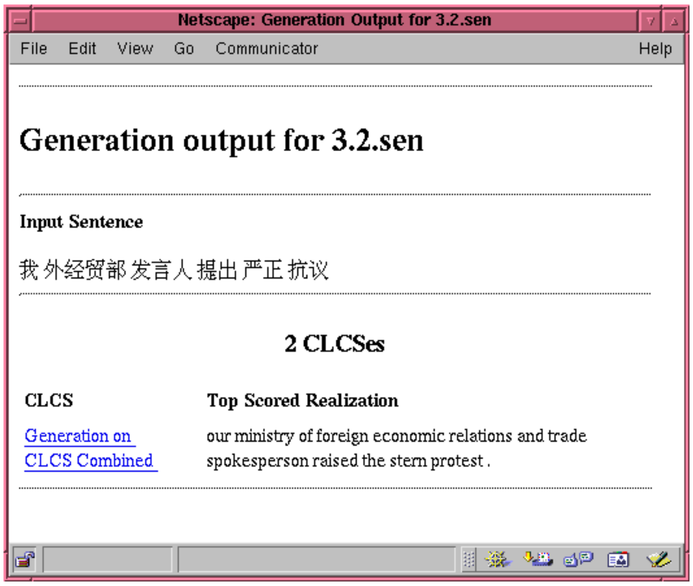

Fig. 2. Translation of a Chinese sentence into English

1. Associate English Optilex glosses with all 12342 Chinese verbs in HowNet, producing 41,324 Chinese-English pairs.

2. Associate each verb-to-concept candidate with at least one of the 500 EVCA classes.[3]

3. For each HowNet concept, partition the associated Chinese-English pairs into groups whose English glosses correspond to EVCA classes.

Fig. 3. Mapping Chinese HowNet Concepts to English EVCA Classes