Cross-Language Frame Semantics Transfer in Bilingual Corpora

R. Basili[1], D. De Cao[1], D. Croce[1], B. Coppola[2], A. Moschitti[2]

[1] Dept. of Computer Science,
University of Roma Tor Vergata, Roma, Italy
{basili,croce,decao}@info.uniroma2.it
[2] University of Trento, Italy
{coppola, moschitti}@disi.unitn.it

Abstract. Recent work on the transfer of semantic information across languages has been recently applied to the development of resources annotated with Frame information for different non-English European languages. These works are based on the assumption that parallel corpora annotated for English can be used to transfer the semantic information to the other target languages. In this paper, a robust method based on a statistical machine translation step augmented with simple rule-based post-processing is presented. It alleviates problems related to preprocessing errors and the complex optimization required by syntax-dependent models of the cross-lingual mapping. Different alignment strategies are here investigated against the Europarl corpus. Results suggest that the quality of the derived annotations is surprisingly good and well suited for training semantic role labeling systems.

1 Motivation

The availability of large scale semantic lexicons, such as Framenet ([1]), has allowed the adoption of a vaste family of learning paradigms in the automation of semantic parsing. Building on the so called *frame* semantic model, the Berkeley FrameNet project [1] has developed a frame-semantic lexicon for the core vocabulary of English since 1997. As defined in [2], a frame is a conceptual structure modeling a prototypical situation. A frame is evoked in texts through the occurrence of its lexical units (LU), i.e. predicate words (verbs, nouns, or adjectives) that linguistically expresses the situation of the frame. Each frame also specifies the participants and properties of the situation it describes, the so called frame elements (FEs), that are the Frame Semantics instantiation of semantic roles. For example the frame CATEGORIZATION has lexical units such as: *categorize,classify,classification,regard*. Semantic roles shared by these predicates, are the COGNIZER (i.e. the person who performs the categorization act), the ITEM construed or treated, the CATEGORY (i.e. the class which the item is considered a member of) and CRITERIA. Semantic Role Labeling (SRL) is the task of automatic labeling individual predicates togheter with their major roles (i.e. frame elements) as they are grammatically realized in input sentences. It has been a popular task since the availability of the PropBank and Framenet annotated corpora [3], the seminal work of [4] and the successful CoNLL evaluation campaigns [5]. Statistical machine learning methods, ranging from joint probabilistic models to support vector machines, have been largely adopted to provide accurate labeling, although inherently dependent on the availability of large scale annotated resources.

It has been observed that the so called resulting *resource scarcity problem* affects a large number of languages for which such annotated corpora are not available [6]. Recent works thus explored