SkyCode MT – a translation system using deep syntactic and semantic analysis

Luchezar Jackov
SkyCode Ltd., Sofia, Bulgaria
PhD student at the Institute for Bulgarian Language,
Bulgarian Academy of Sciences
lucho@skycode.com

Abstract

SkyCode MT is a rule-based machine translation system that evaluates all possible parsing hypotheses and ranks them using dependency relations. It uses Princeton WordNet (PWN) (Fellbaum, 1998) synsets as universal dictionary and has separate per-language analysis and synthesis modules which enables translation between any two of the seven languages of the system. It has been developed as a complete solution used in commercial applications. The small footprint allows its use on mobile devices (smartphones and tablets). The system has participated as a translation vendor in the 7th FP project iTranslate 4 (http://itranslate4.eu).

1 System description

The system translates between English, German, French, Spanish, Italian, Turkish and Bulgarian by means of a deep internal syntactic and semantic representation of the input text. This allows the translation of the 21 language pairs (42 translation directions) in just 150 MB. The sense inventory is based on the original PWN synonym sets (concepts) extended with lexicalizations having the following synset coverage: 74124 in Bulgarian, 62015 in Turkish, 79553 in German, 84345 in Spanish, 88955 in French and 78718 in Italian.

The lexicalizations are used for morphological analysis of the source, creating initial hypotheses for simple concepts (the various readings of single words and collocations). The system uses manually defined rules to generate all possible parses (parsing hypotheses) for the source by applying them in a bottom-up fashion on adjacent hypotheses, building an entire sentence parse tree. The rules are based on Chomsky-normal-form context-free grammar extended with dependency relations on the constituents. As a result each hypothesis identifies concepts (PWN synsets) and dependency relations between them. The relations between the concepts are used for evaluation of how 'sensible' the hypothesis is by consulting a relations knowledge base. It is defined on the PWN synsets and is language-independent for most of the relations.

The translation is synthesized using the PWN synset lexicalizations for the target language and manually defined synthesis rules, transferring the semantic relations to the translation.

Both the synthesis and the analysis rules are shared between languages that have common linguistic phenomena such as the same word order, e.g. S → NP VP, VP → V NP, VP → V PP.

The use of PWN synsets as universal dictionary and knowledge base as well as splitting the analysis from the synthesis allow for the translation between the languages of the system without having to define per-language-pair rules. This also makes adding a new language relatively easy by only defining PWN lexicalizations, and analysis and synthesis rules specific to the new language.

The system is implemented in C++, which makes it portable across various operating systems and platforms including mobile devices. A detailed description is given in (Jackov, 2014).

References

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Jackov, L. 2014. *Machine translation based on WordNet and dependency relations.* In Computer Linguistics In Bulgaria 2014, p. 64−72.

© 2017 The authors. This article is licensed under a Creative Commons 3.0 licence, no derivative works, attribution, CC-BY-ND.