Introduction

Collocation measures are applied to large parsed corpora (Lehmann and Schneider, 2012), followed by manual filtering to automatically extract several categories of multi-word entities (MWE). Our dependency parser Pro3Gres (Schneider, 2008) combines a hand-written competence grammar, which represents the syntax principle, with statistical performance disambiguation, which represents the idiom principle (Sinclair, 1991).

Verb-Preposition Structures

We use O/E as collocation measure. O/E has a tendency to report rare collocations: in traditional window-based approaches, garbage appears at the top.

- Approaches based on parsed corpora provide considerably cleaner data (Seretan, 2011)
- Paired with a T-score significance threshold O/E delivers very good results.
- 2nd key criterion is fixedness. We use Yule’s k as a measure of diversity: proven independence on token counts.

Light Verb Constructions

Our automatic detection of light verbs is described in Ronan and Schneider (submitted). We use several collocation measures.

Evaluation: give Precision & Recall on BNC, using T-Score & simple filter

Subject-Verb-Object and Others

Strong but gradient idiomatic and selectional preferences prevail on all levels, e.g. verb-object, subject-verb, in syntactic structures, morphology, alternation preferences.

In addition to extracting MWE classes with arbitrary borders, abstracting to probabilistic interdependent features is useful:

- Bi-lexical preferences (Collins, 1999; Hoey, 2005)
- Construction grammar (Stefanowitsch and Gries, 2003)
- Information-theoretic measures such as surprisal (Levy and Jaeger, 2007)

Exercising Priming is the key factor for Hoey (2005): “lexis is complexly and systematically structured and that grammar is an outcome of this lexical structure” (1).

“We can only account for collocation if we assume that every word is mentally primed for collocational use” (8)

Pawley and Syder (1983, 193): “native speakers know best how to play the game of fixedness vs. expressiveness: “native speakers do not exercise the creative potential of syntactic rules to anything like their full extent, and that, indeed, if they did so they would not be accepted as exhibiting nativelike control of the language.”

Levy and Jaeger (2007): “UID (uniform information density) can be seen as minimizing comprehension difficulty”.

Language learners produce less fixed, less entrenched structures. We use the NICT Japanese Learner English (JLE) Corpus. It contains 120,000 sentence pairs consisting of an original language learner sentence and a corrected sentence. Bigram surface surprisal log 1/ P(wi) + log 1/ P(wi | wi-1) has a mean of 11.7 (and SD=3.36) for corrected text, and 11.5 (and SD=3.48) for original learner text. Comparison:

Parser as Model of Fixedness

We also apply the parser to Learner English. We have manually annotated 100 sentence pairs from the NICT Japanese Learner English (JLE) Corpus. A parser is a language model because:

- It takes attachment decisions (predictions) based on grammar rules and lexical preferences
- It learns form real-word data: syntactically annotated Penn treebank
- Fitting the model: Entrained structures get higher scores, as they are expected. L2 utterances do not fit the model very well. They abide less to priming, contain more information in Shannon’s terms.

For the investigation of highly gradient, complex and interacting factors a parser-based language model is useful. We show that:

- Parser performance is significantly lower for the original Learner data than for the corrected (see Figure 1);
- Parser scores are significantly lower for the original Learner data than for the corrected (see Figure 2).

Subject-Verb-Object and Others

Parsers aggregating the probabilistic information from all levels

Subject-Verb-Object and Others

- Parsers aggregating the probabilistic information from all levels

References


