Gaillard, Benoît; Navarro, Emmanuel; Gaume, Bruno
CLLE-ERSS, Univ. Toulouse 2, benoit.gaillard@univ-tlse2.fr
IRIT, Univ. Toulouse 3, navarro@irit.fr
CLLE-ERSS, Univ. Toulouse 2, bruno.gaume@univ-tlse2.fr

From binary synonymy to near synonymy by optimal proxemy of lexical resources

Dictionaries of synonyms are often encoded as binary links between words. This encoding raises many issues. For example, we show that there is very little agreement between different expert-built resources, whereas they represent a similar linguistic reality. This may be due to the fact that absolute synonyms are rare and that most of the synonymy described in the dictionaries is in fact a description of nuances of a general meaning shared by a set of near-synonyms (Edmonds and Hirst, 2002). The complexity and scale of these nuances leave room for interpretation when projecting them on the basic "synonym/not synonym" alternative, which in turn gives rise to discrepancies between electronic lexical resources at the most fine-grained level. However, we claim that, at a coarser level, the patterns drawn by sets of near-synonyms should be mostly independent from the resource used to describe the lexicon of a given language. We propose a model that leverages the binary encoding of resources to represent the synonymy structure at various levels of granularity. This is not a new merging method, but a attempt at a better modelling of the notion of synonymy. Based on this model, we find an optimum level of granularity for which resources are the most similar, therefore our approach seems appropriate in order to infer near-synonymy patterns from any particular binary encoded resource.

We studied the similarities of seven well-known french dictionnaries, binarily encoded as graphs in which vertices are words and two vertices are linked by an edge if they are considered synonyms. We used the F-score between sets of edges to measure the similarities between the unweighted graph-encoded dictionaries of synonyms and found that it does not exceed 0.5. To model these resources on coarser levels, we introduce a notion of *proxemy* between any two words, based on the probability of reaching one word from another after a random exploration of the graph. At each step of the exploration, a particle is modelled to move from one word to any of its neighbours with a probability inversely proportional to its degree. Each word is thus associated to a *proxemy vector* characterising its semantic similarity with the other words. This vector depends on the starting vertex and on the number of exploration steps. When the number of steps tends to infinity, all words have the same proxemy vector, which describes the importance of each word in the graph, at the highest level of generality (Gaume et al., 2010). Conversely, after only one step, proxemy vectors have the lowest level of generality, taking only the neighbours of each node into account. The distance between two resources is then defined as the mean euclidian distance between the proxemy vectors of the same word in the two graphs. We experimentally found an optimal number of proxemy steps for which the similarity of any two resources is maximal. This is an optimal level of generality, according to the proxemy model, that enables us to discover near-synonymy patterns that are similar across dictionaries whose binary synonymy patterns are very different.

Bibliography:

P. Edmonds and G. Hirst (2002) *Near-synonymy and lexical choice*, Computational Linguistics, 28(2), pp. 105—144.

B. Gaume, F. Mathieu and E. Navarro (2010) *Building Real-World Complex Networks by Wandering on Random Graphs*, Information - Interaction – Intelligence, (To appear).