Maja Miličević Petrović*, Radoslava Trnavac, Borko Kovačević****

*University of Bologna, **University of Belgrade

maja.milicevic2@unibo.it, radoslava.trnavac@fil.bg.ac.rs, borko.kovacevic@fil.bg.ac.rs

Can linguistic insights on semantic similarity help its automatic treatment?

Keywords: cross-level semantic similarity, paraphrase, taxonomy, natural language processing

Establishing semantic similarity, i.e. whether and to which degree the meanings of two text items are similar to each other, is a central natural language processing task, and an integral part of several more complex tasks, such as information retrieval. The models used in automatic treatment of semantic similarity rely on different types of information, including linguistic features. However, these features tend to be rather basic, with automatically assigned syntactic dependencies being among the most elaborate ones. This is partly due to the tendency of contemporary NLP models to be based on non-transparent statistical representations (e.g. word embeddings), but it is also related to the fact that semantic similarity is not systematically defined and described in linguistics, despite being relevant for most levels of analysis – lexical semantics to begin with, but also morphology and syntax (e.g. in the study of diathesis alternations; Vila et al. 2014).

In this paper, we propose a taxonomy of semantic similarity types and indicators, based on the classifications of paraphrase previously proposed by Vila Rigat (2012), Vila et al. (2014), Milićević (2007) and Mel'čuk (2012), with a focus on the nature of information that similarity is based on. The detection of paraphrase, intended as the relationship between linguistic expressions with different wording and (approximately) the same meaning, is a task closely related to semantic similarity – similar units can be seen as paraphrases of each other. Another relevant overlap concerns the fact that a relation of either paraphrase or similarity can be established between units of different size, such as a phrase and a sentence (see Mel'čuk 2012: 46). The main differences between the two phenomena concern the fact that semantic similarity is a more general concept, and while paraphrase tends to be treated in binary (yes/no) terms in NLP tasks (Vila Rigat 2012), semantic similarity is more commonly identified and annotated through finer-grained degrees (typically using Likert scales, see Jurgens et al. 2014).

The proposed taxonomy is shown in Table 1. A core distinction it implements is that between linguistic, quasi-linguistic and extralinguistic similarity. Linguistic similarity is (primarily) based on language-internal information, and has multiple subtypes. The quasi-linguistic domain captures inference-based similarity that relies on pragmatic information, while the extralinguistic domain involves information equivalence, but requires knowledge external to language to be used. To empirically validate the proposed taxonomy, we are currently creating the *CLSS.news.sr-ling* dataset, comprising 1,000 phrase-sentence and 1,000 sentence-paragraph newswire text pairs in Serbian, based on the *CLSS.news.sr* corpus (https://vukbatanovic.github.io/CLSS.news.sr/), already annotated with cross-level semantic similarity scores (on a scale 0-4), to which information about the semantic similarity categories is being added. We will discuss the distribution of different semantic similarity types and indicators in the dataset, and propose several quantitative measures based on them, explaining how this kind of information can be taken into account in NLP models.

Word count: 467

References

Jurgens, D., M. T. Pilehvar and R. Navigli (2014). SemEval-2014 Task 3: Cross-Level Semantic Similarity. *Proceedings of SemEval 2014*. 17-26.

Mel’čuk, I. A. (2012). *Semantics. From Meaning to Text*. Amsterdam: John Benjamins.

Milićević, J. (2007). *La paraphrase*. Bern: Peter Lang.

Vila Rigat, M. (2013). *Paraphrase Scope and Typology. A Data-Driven Approach from Computational Linguistics*. PhD dissertation, University of Barcelona.

Vila, M., A. Martí and H. Rodríguez (2014). Is this a paraphrase? What kind? Paraphrase boundaries and typology. *Open Journal of Modern Linguistics* 4. 205-218.

Table 1. Overview of the taxonomy of semantic similarity

Similarity type	**Indicator type**	**Indicator subtype**	**Indicator**
Linguistic	Morpholexicon-based	Morphology-based	Identical Inflectional Derivational
		Lexicon-based	Spelling and format Same polarity (synonymy, hyponymy, meronymy) Synthetic/analytic Opposite polarity Converse
	Structure-based	Syntax-based	Diathesis alternations Negation switching Ellipsis Coordination changes Subordination and nesting changes
		Discourse-based	Punctuation Direct/indirect style Sentence modality
	Semantics-based		
	Miscellaneous		Change of format Change of order Addition/deletion
Quasi-linguistic	Pragmatic		
Extralinguistic	Situational		
	Encyclopaedic		
	Logical		