

Detecting and Processing Figurative Language in Discourse

Caroline Sporleder

Universität Trier

sporledc@uni-trier.de

ABSTRACT

Figurative language poses a serious challenge to NLP systems. The use of idiomatic and metaphoric expressions is not only extremely widespread in natural language; many figurative expressions, in particular idioms, also behave idiosyncratically. These idiosyncrasies are not restricted to a non-compositional meaning but often also extend to syntactic properties, selectional preferences etc. To deal appropriately with such expressions, NLP tools need to detect figurative language and assign the correct analyses to non-literal expressions. While there has been quite a bit of work on determining the general ‘idiomaticity’ of an expression (type-based approaches), this only solves part of the problem as many expressions, such as *break the ice* or *play with fire*, can also have a literal, perfectly compositional meaning (e.g. *break the ice on the duck pond*). Such expressions have to be disambiguated in context (token-based approaches). Token-based approaches have received increased attention recently. In this talk, I will present an unsupervised method for token-based idiom detection. The method exploits the fact that well-formed texts exhibit lexical cohesion, i.e. words are semantically related to other words in the context. I will show how cohesion can be modelled and how the cohesive structure can be used to distinguish literal and idiomatic usages and even detect newly coined figurative expressions.

KEYWORDS: Discourse, Figurative Language, Token-Based Idiom Detection.
