A Short Note on Social-Semiotic Networks from the Point of View of Quantitative Semantics

Alexander Mehler

Goethe-University Frankfurt am Main
Mehler@em.uni-frankfurt.de

Summary. In this extended abstract we discuss four related characteristics of semantic spaces as the standard model of meaning representation in quantitative semantics. We argue that these characteristics are challenged from the point of view of social web communities and the possibilities which they offer in terms of exploring semantic *and* pragmatic data. More specifically, we plead for a reconstruction of the weak contextual hypothesis in order to account for non-linguistic, pragmatic aspects of context. Finally, we mention two consequences of such a pragmatic turn, that is, in the area of named entity recognition and of language evolution.

Predominant models of quantitative semantics are based on assumptions which are challenged from the point of view of the possibilities offered by exploring social web communities. More specifically, approaches to usage-based geometrical models of meaning in the form of semantic spaces make — more or less explicitly — use of the following assumptions:

- *Pragmatic homogeneity:* In order to learn the usage regularities of a word, for example, one starts from a corpus of texts which is assumed to be pragmatically homogeneous in the sense that the variety of authors, genres (functions or purposes), locations and times of origin of these texts has no impact on their status as equally reliable resources of exploring the meaning of that word. In other words: Each occurrence of a word is seen to be equally reliable to contribute to its usage regularities irrespective of the pragmatic context of that co-occurrence.
- *Context as cotext:* As a result of (inherently) assuming this kind of pragmatic homogeneity, the *context*-sensitivity of sign meaning (Barwise and Perry, 1983) is analyzed from the point of view of *cotext*-sensitivity only. That is, context building units are solely accounted for in terms of linguistic units surrounding the occurrences under consideration — disregarding the pragmatic context according to which the same word has different meanings in different contexts. Recent trends in corpus linguistics overcome this pragmatic insufficiency. However, when analyzing the formal

Dagstuhl Seminar Proceedings 08391
Social Web Communities
http://drops.dagstuhl.de/opus/volltexte/2008/1788

ingredients of predominant models of semantic spaces one searches vainly for a representation model of *non*-linguistic context.

- *Learnability:* Related to these two assumptions is a predominant procedural model where the meanings of signs are synchronously learnt as meaning points in semantic space. Obviously, this approach does not reflect that the meanings of signs are learnt asynchronously — within a speech community as well as by a single member of that community. Semantic spaces classically start from an input corpus of fixed size which is processed as a whole by being mapped onto a term-document matrix as input to some mathematical operations (including, e.g., single value decompositions or cosine measurements — cf. Berry and Browne 2005). As a consequence, dynamically growing corpora whose extent is initially unknown are hardly taken into consideration when it comes to building semantic spaces.[1] However, this is what human beings do when learning lexical meanings: processing discourses in an iterated manner but not all at once.
- *Single agent-based learning:* To finish this short analysis of assumptions underlying approaches to semantic spaces let us finally mention that the latter ignorance of iterative learnability relates to the fact that semantic spaces as built, e.g., by *Latent Semantic Analysis* (LSA) (Landauer and Dumais, 1997) instantiate the class of single agent models. That is, the learning algorithm behaves as a "lonely agent" who processes all its input texts in isolation irrespective of any communication with other agents of the same community. However, meanings of lexical units are shared among members of a speech community since they are learnt in a process of distributed cognition (Steels, 1996; Hollan et al., 2000; Christiansen and Kirby, 2003; Kirby, 2002) and, therefore, they are represented in a distributed manner among the agents of that community (let alone the propagation of linguistic knowledge among consecutive generations — cf. Kirby and Hurford 2002).

It turns out that we are now in a position in which we can explore data which because of its quality *and* quantity may help to overcome these shortcomings of semantic spaces.

Let us analyze this potential with a focus on the weak contextual hypothesis of Miller and Charles (1991). It says that the similarity of the contextual representations of words contributes to their semantic similarity. In terms of explorative, emergent semantics we can reformulate this by saying that the semantic similarity of signs is a function of the similarity of the contexts in which they occur. Obviously, this approach goes beyond simply counting *co*-occurrences within the same contexts — such an approach is restricted to exploring syntagmatic associations among signs and, therefore, hardly deals with their paradigmatic relations (Raible, 1981). In contrast to this, the weak

[1] Think, for example, of the Wikipedia which may be input to building semantic spaces which evolve according to the insertions, deletions and modifications which users make to Wikipedia articles.

contextual hypothesis accounts for paradigmatic associations by exploring the similarity of the contexts in which the signs occur. However, as already stated above, the notion of context is instantiated in terms of linguistic units, e.g. sentences, paragraphs, texts or web pages, that is, in terms of *co-texts* (as a sort of purely *linguistic* contexts) — cf. Widdows (2003).

Obviously, this focus on cotext is deficient as it does not sufficiently account for the semantic diversification and the semantic dynamics of signs as a result of the pragmatic diversity of their use. With the rise of social web communities we are in a position that we can leave this narrow focus of quantitative semantics. That is, when analyzing social web communities we do not only have access to the linguistic manifestations of the interactions of agents (e.g. in terms of tags, collectively written wiki articles etc.). In fact, we also have access to manifestations of the participants of these interactions (e.g. in terms of personal profiles manifesting their social role within the corresponding community) and related pragmatic variables. Amongst others, this includes the time and duration of the interactions among agents which are partly accessible by the history function of wikis and related information systems. As a consequence of this qualitative enrichment of accessible pragmatic information we can think of a pragmatically enriched reformulation of the weak contextual hypothesis. This might look as follows:

> The similarity of the pragmatic contexts of the uses of words contributes to their semantic similarity.

An approach to semantic spaces in the line of this extended notion of context may help to open the door for a *pragmatic turn* of quantitative semantics. Take the example of synonymy. We may qualify the question for the synonymy of two words by distinguishing the group of agents, their communicative purposes or the period of time in which these words are partly synonymous. Likewise, we do not need to represent the meaning of a polysemous word as a one-to-many relation between that word and its different readings. In contrast to this, we can think of a relational model — very alike to situation semantics but on the grounds of a geometric or at least topological model (Rieger, 2002; Gärdenfors, 2000) — by which we additionally represent pragmatic variables. That way, any statement about the semantic similarity or relatedness of signs is specified in terms of groups of agents who have generated these signs in certain contexts to meet certain functions and purposes of their communication. In many cases of social web communities there is much more information available about the pragmatic conditions of communication, much more than has been accessible to purely text-based corpus analyzes at any time before in computational linguistics. As a result of such an approach we get an understanding of semantic spaces which are no longer seen to be homogeneous in the sense that information about the underlying agent community is abstracted away — as if a single abstract agent would have produced the text corpus underlying the build-up of the semantic

space. In fact, we may build stratified semantic spaces which split into different subspaces each related to a certain community and probably overlapping in order to enable communication across-the-board of communities.

One can think of many consequences of this pragmatic turn. We mention only two of them:

- *Networked entity recognition:* In contrast to predominant approaches to named entity recognition we may reconsider structuralist models of object identity according to which entities are specified by their relative position in a network of (sign) relations — in our case a network of a social web community. However, any such pragmatic grounding of networked entity recognition asks for distinguishing between semiotic networks (whose vertices are signs and whose edges stand for certain relations of these signs) on the one hand and the social community which by virtue of its web-based communication brings about the latter semiotic network thereby being shaped and stabilized by this very network. In other words: when dealing with semiotic networks one has to deal with the underlying social (agent) networks and vice versa so that semantics gets strongly connected to pragmatics. Such multi-level networks (which because of their internal structure go beyond simple n-partite graphs) can be called *social-semiotic networks*.
- *Forecasting linguistic dynamics:* Social-semiotic networks, their topology and dynamics are the very object of complex network analyzes of the semantics of web-based units. It can be made an object of the rising area of simulation models of language evolution. More specifically, we can ask about the mutual impact of networking on the social level on the one hand and on the semiotic level on the other. See, for example, Gong and Wang (2005) who consider the community as the dependent variable whose structuring occurs dependent on structure formation on the semiotic level. In contrast to this, Mehler (2008) looks on structure formation on the semiotic level thereby exploring the community model as the independent variable. What may be done now is to combine these two models in order to study the dynamics of community building and semiotic network formation in a single model thereby integrating the naming with the association game. It turns out that social web communities provide the data by which such simulation models can be empirically tested.

Obviously, we face the merging of two related areas: quantitative semantics on the one hand and simulation models of language evolution on the other. That this merging is a realistic stage of the development of quantitative semantics is due to the quantity of high-quality semantic and pragmatic data being available by the traces which emergent social web communities leave behind.

References

Barwise, J. and Perry, J. (1983). *Situations and Attitudes*. MIT Press, Cambridge.

Berry, M. W. and Browne, M. (2005). *Understanding Search Engines. Mathematical Modeling and Text Retrieval*. SIAM, Philadelphia.

Christiansen, M. H. and Kirby, S. (2003). Language evolution: Consensus and controversies. *Trends in Cognitive Sciences*, 7(7):300–307.

Gärdenfors, P. (2000). *Conceptual Spaces*. MIT Press, Cambridge, MA.

Gong, T. and Wang, W. S.-Y. (2005). Computational modeling on language emergence: A coevolution model of lexicon, syntax and social structure. *Language and Linguistics*, 6(1):1–41.

Hofmann, T. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196.

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transaction on Computer-Human Interaction*, 7(2):174–196.

Kirby, S. (2002). Natural language from artificial life. *Artificial Life*, 8(2):185–215.

Kirby, S. and Hurford, J. R. (2002). The emergence of linguistic structure: An overview of the iterated learning model. In Cangelosi, A. and Parisi, D., editors, *Simulating the Evolution of Language*, chapter 6, pages 121–148. Springer, London.

Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Leopold, E. (2005). On semantic spaces. *LDV Forum*, 20(1):63–86.

Mehler, A. (2007). Compositionality in quantitative semantics. A theoretical perspective on text mining. In Mehler, A. and Köhler, R., editors, *Aspects of Automatic Text Analysis*, Studies in Fuzziness and Soft Computing, pages 139–167. Springer, Berlin/New York.

Mehler, A. (2008). On the impact of community structure on self-organizing lexical networks. In Smith, A. D. M., Smith, K., and Ferrer i Cancho, R., editors, *Proceedings of the 7th Evolution of Language Conference (Evolang 2008), March 11-15, 2008, Barcelona*, pages 227–234. World Scientific.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1):1–28.

Raible, W. (1981). Von der Allgegenwart des Gegensinns (und einiger anderer Relationen). Strategien zur Einordnung semantischer Informationen. *Zeitschrift für romanische Philologie*, 97(1-2):1–40.

Rieger, B. B. (2002). Semiotic cognitive information processing: Learning to understand discourse. A systemic model of meaning constitution. In Kühn, R., Menzel, R., Menzel, W., Ratsch, U., Richter, M. M., and Stamatescu, I. O., editors, *Perspectives on Adaptivity and Learning*, pages 347–403. Springer, Berlin.

Schütze, H. (1997). *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*, volume 71 of *CSLI Lecture Notes*. CSLI Publications, Stanford.

Steels, L. (1996). Self-organising vocabularies. In Langton, C. G. and Shimohara, K., editors, *Proceedings of Artificial Life V, Nara, Japan*, pages 179–184.

Steels, L. (2000). The puzzle of language evolution. *Kognitionswissenschaft*, 8:143–150.

Steels, L. (2006). Collaborative tagging as distributed cognition. *Pragmatics & Cognition*, 14(2):287–292.

Widdows, D. (2003). A mathematical model for context and word-meaning. In *Fourth International and Interdisciplinary Conference on Modeling and Using Context, Stanford, California, June 23-25*, pages 369–382.

Widdows, D. (2004). *Geometry and Meaning*. CSLI Publications, Stanford.