Towards the Realistic Natural Language Representations

Petr Sojka

Faculty of Informatics, Masaryk University
Botanická 68a, 602 00 Brno, Czech Republic
sojka@fi.muni.cz

Abstract. This essay suggests a way to derive a natural language representation from textual corpora into the connectionist, continuous representations. Based on the lexical priming theory and psycholinguistic evidence we discuss benefits and potential of alternative representations inspired by connectionist approaches towards *computation* of personalized mental lexicon from and during empirical language usage.

Keywords: natural language representation, priming, lexical priming, semantic priming, data discretization, language modelling, representation of meaning, personal mental lexicon, empirical linguistics

On resiste a l'invasion des armees; on ne resiste pas a l'invasion des idees.
(Victor Hugo (1802–1885))

1 Striving to Getting an Insight

Linguists try to get insight into language communication by naming the language phenomena. They named morphology, syntax, semantics and pragmatics as levels of natural language processing and understanding, usually hoping that solving lower level is a prerequisite to tackle the higher one. They pose questions like "Do Word Meaning Exist?" [6]. They try to embrace the *knowledge of a language* as a set of grammars, dictionaries and the battery of rules to model the forms of communication via natural language.

Psycholinguistics is concerned with the ability of human brain to understand and generate language. It tries to understand the cognitive processes that make it possible *to communicate the thoughts and knowledge* via language. The recent research on associative, semantic and thematic *priming* effects [10] shows evidence that language lexicalization plays irreplaceable rôle in the *conceptual organization of knowledge*.

Computational linguists design algorithms to verify or deny theoretical linguists' theories usually by modelling the language usage from their surface form. They often use big corpora to build a language model based on statistics computed from texts by zillions of writers. The natural language representation is based on averaging of word usage. The *representation of knowledge of a language* is stored in the form of big corpora containing billions of words [16].

Computer scientists are trying to understand the computation by designing appropriate data structures that allow appropriate representation of the problem in hand, so that algorithms are easily formulated. They have recently came up with a new definition of computation as any *process generating knowledge* [14]. It fits the view of natural language understanding as a computational process, during which word meanings and semantics gets computed on the fly during discourse, and the representation might be affected by such computation.

All research communities above strive to name, generate and compute knowledge of natural language understanding to get an insight on how to model natural language communication. The common sense is that the key to successful natural language processing is appropriate *natural language representation* (NLR).

You shall know a word by the company it keeps. (John Rupert Firth)

2 Development of Natural Language Representations

Chomskyan linguistic nativism [3] stressed the generative, formal qualities of language, ignoring the fact that people communicate successfully even using syntactically wrong discourse. Similarly, on a semantic level, the twentieth century prevalent view was that a word does objectively have several distinct *meanings* that could be enumerated as in a dictionary entry. The claim was that by solving the task of word sense disambiguation we will be close to natural language understanding. Just another example of another *discrete representation* in language modelling is represented by the view that some powerful logic will be sufficient to effectively represent discourse semantics.

Corpora linguists collected the evidence that language use is very variational and diverse, not fitting the boundaries of syntactic, grammatical, semantic structures and logical formalisms. Language is on move, with many irregularities that develop in time and space. By studying *word sketches* [12] we see that word meanings are subjective, hard to separate, and form collocates depending on context. Since the end of last millenium, there are linguists that do not believe in clear separation of word senses [11].

To anchor a word in a context, the theory of *lexical priming* has been coined by Hoey [8]. Backed up by evidence from psycholinguistics, he articulates and argues for a new theory where each occurrence of lexical item enforces 'priming' of it given a co-*locational* context. A Firth's 'word's company' is viewed broadly, as pervasive and subversive types of collocations on a sentence and higher levels. The word, or more precisely a lexical item, is learnt through encounters. Each "new encounter either reinforces the priming or loosens it," and make "drifts in the priming" [8].

Lexical priming theory is convincing in many aspects, especially that it allow the explanation of how different word meanings may come up based on previous word usage, and the whole context of lexical occurrence including the pragmatics. The contextual clues additively contribute to the on the fly computation of every word meaning. This is in sync with all the WSD research to date. Natural question arises: how to implement the *computational lexical*

priming for use in NLP tasks, and what representations should be used? Could word sense disambiguation problem be solved by lexical priming computed over appropriate data structures mimicking the way of processing we collect evidence from psycholinguistic research and novel view of computation?

> We should stop acting as if our goal is to author extremely elegant theories, and instead embrace complexity and make use of the best ally we have: the unreasonable effectiveness of data. (Peter Norvig *et al.*, 2009 [5])

3 Discrete and Continuous Language Representations

Most mainstream "scholastic" language representations used to date are *discrete* representations as lists, graphs or logics, and aims to capture a language as an objective, discrete, fossil structure. It does work to some extent for modelling of *conscious*, deductive reasoning, but leads to a very limited functionality and use cases.

In the real world, language nuances of every communication side are different, and it takes time before a word's usage and meaning will settle, converge into an entry in dictionary and will be understood during the man to man discourse. Workflow of language processing is usually layered into separate modules of morphology, syntax, and semantics, forgetting the primary communication goal of discourse via natural language. Syntax encodes information structure [2], only helps to resolve the main task of meaning disambiguation of a message.

One should be warned by adopting easy simplifications. "A linear ordering of a multi-parameter universe is usually nonsense" [15] does hold not only in an science impact measurements, but in the word meaning, or generally, in language modelling, too. A more complex representation is necessary.

> . . . if nature is really structured with a mathematical language and mathematics invented by man can manage to understand it, this demonstrates something extraordinary. The objective structure of the universe and the intellectual structure of the human being coincide. (Pope Benedict XVI [1])

4 The Unreasonable Effectiveness of Language Representations Computed from Corpora

Let us suppose that lexical items (single word, lemma or longer term) will be represented as a node in a neural network. Let synapses represent co-locative relations of different kinds, including perceptual clues from visual subsystem, trains of thoughts, coherence links between sentences etc.

Mutatis mutandis, methods like Hebbian learning [7], WebSOM [13] and random walking in graphs (explicit collocation representations) may be used in the computations of continuous representations of natural language.

There are well established methods of building a language models by various types of *smoothing*. Failure of 'semantic web' approaches to unify

(discrete) keyword-based and ontology based semantics cause shifting towards (continuous) distributional semantics approaches.

There is an evidence of *conscious and unconscious* semantic priming [4]. Corresponding *discrete and continuous* data structures might help to proper modelling of personal mental lexicon. We are developing appropriate discretization algorithms specific for natural language tasks, based on random walking [9] in huge corpora towards this goal.

Get comfortable with paradoxes. (David Allen)

5 Conclusion

We have expressed our view of continuous and personal language representation motivated by Hoey's lexical priming theory. We argue that it will allow modelling of several language phenomena with ease. It is yet to be confirmed by computational experiments and computed representations appropriate for specific NLP tasks in natural language understanding.

Acknowledgements This work has been partially supported by the European Union through its Competitiveness and Innovation Programme (Information and Communications Technologies Policy Support Programme, "Open access to scientific information", Grant Agreement No. 250,503).

References

1. Message of His Holiness Benedict XVI to Archbishop Rino Fisichella, Rector Magnificent of the Pontifical Lateran University, on the Occasion of the International Conference "From Galileo's Telescope to Evolutionary Cosmology. Science, Philosophy and Theology in Dialogue" (Nov 2009)
2. Brown, M., Savova, V., Gibson, E.: Syntax encodes information structure: Evidence from on-line reading comprehension. Journal of Memory and Language 66(1), 194–209 (2012), http://dx.doi.org/10.1016/j.jml.2011.08.006
3. Chomsky, N.: Syntactic Structures. Walter de Gruyter (2002)
4. Dehaene, S., Naccache, L.: Imaging unconscious semantic priming. Nature 395(6702), 597 (1998)
5. Halevy, A., Norvig, P., Pereira, F.: The unreasonable effectiveness of data. IEEE Intelligent Systems 24(2), 8–12 (Mar 2009), http://dx.doi.org/10.1109/MIS.2009.36
6. Hanks, P.: Do word meanings exist? Computers and the Humanities 34, 205–215 (Apr 2000)
7. Hebb, D.: The Organization of Behavior. Wiley, New York, second edn. (1968)
8. Hoey, M.: Lexical Priming: A New Theory of Words and Language. Routledge (2012)
9. Hughes, T., Ramage, D.: Lexical semantic relatedness with random graph walks. In: Proceedings of EMNLP-CoNLLi 2007. pp. 581–589 (2007)
10. Jones, L.L., Estes, Z.: Lexical priming. Visual Word Recognition Volume 2: Meaning and Context, Individuals and Development 2, 44 (2012)

11. Kilgarriff, A.: I don't believe in Word Senses. Computers and the Humanities 31(2), 91–113 (1997)
12. Kilgarriff, A., Tugwell, D.: WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In: Proceedings of the Workshop 'Collocation: Computational Extraction, Analysis and Exploitation' ACL, Toulouse, France. University of Brighton (2001), http://www.itri.bton.ac.uk/~{}David.Tugwell/colloc.ps
13. Lagus, K., Kaski, S., Kohonen, T.: Mining massive document collections by the websom method. Inf. Sci. 163(1–3), 135–156 (Jun 2004), http://dx.doi.org/10.1016/j.ins.2003.03.017
14. van Leeuwen, J., Wiedermann, J.: Computation as an unbounded process. Theoretical Computer Science 429, 202–212 (2012), http://dx.doi.org/10.1016/j.tcs.2011.12.040
15. Maurer, H.: A linear ordering of a multi-parameter universe is usually nonsense. Theoretical Computer Science 429, 222–226 (2012)
16. Pomikálek, J., Jakubíček, M., Rychlý, P.: Building a 70 billion word corpus of English from ClueWeb. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC '12). European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)