Statistical Language Modeling

Universität des Saarlandes

La théorie des probabilités n'est, au fond, que le bon sens réduit au calcul.

Probability theory is nothing but common sense reduced to calculation.

–Pierre-Simon Laplace, 1812

Winter Semester 2014/2015
Mondays 12:00–14:00, Starting 27 Oct. 2014
Place: 2.11 Geb. C7.2

Instructor	Dr. Jon Dehdari
Office	**DFKI**: 1.11 Gebäude D3.1; **FR4.6**: 1.15 Gebäude A2.2
Office Hours	By appointment
Email	`jon.dehdari` at `dfki.de`
Website	`jon.dehdari.org/teaching/uds/lm`

1 Overview / Überblick

Statistical language modeling, which provides probabilities for linguistic utterances, is a vital component in machine translation, automatic speech recognition, information retrieval, and many other language technologies. In this seminar we will discuss different types of language models and how they are used in various applications. The language models include *n*-gram- (including various smoothing techniques), skip-, class-, factored-, topic-, maxent-, and neural-network-based approaches. We also will look at how these perform on different language typologies and at different scales of training set sizes.

This seminar will be followed at the end of the Winter Semester by a short project (seminar) where you will work in small groups to identify a shortcoming of an existing language model, make a novel modification to overcome the shortcoming, compare experimental results of your new method with the existing baseline, and discuss the results. It'll be fun.

2 *Tentative* Schedule / *Vorläufiger* Terminplan

Date Datum	Topic Thema	Presenter(in)	Q's Fragen
Mo, 20 Oct.	Formalities & Topic Assignment Formalitäten u. Themenvergabe		
Mo, 27 Oct.	*n*-gram LMs: Koehn (2010, ch 3 & pp. 181–198)		
Mo, 3 Nov.	**No Class / Kein Seminar**; do assignment 1		
Mo, 10 Nov.	Witten-Bell, Kneser-Ney, MKN: Koehn (2010, pp 199–203)		
Mo, 17 Nov.	Cache: Kuhn and De Mori (1990) & Skip: Huang *et al.* (1993, §6.1) Goodman (2001, §4) (Guthrie *et al.*, 2006)		
Mo, 24 Nov.	Factored LMs: Bilmes and Kirchhoff (2003)		
Mo, 1 Dec.	Sentence Mixture LMs: Iyer and Ostendorf (1999) (Goodman, 2001, §7)		
Mo, 8 Dec.	Topic LMs: Gildea and Hofmann (1999) (Tan *et al.*, 2011; Bellegarda, 2000; Blei *et al.*, 2003)		
Mo, 15 Dec.	Class-based & Model M: (Brown *et al.*, 1992) Chen (2009)		
Mo, 5 Jan.	FeedForward NN: Bengio *et al.* (2003)		
Mo, 12 Jan.	Recurrent NN: Elman (1990); Mikolov *et al.* (2010)		
Mo, 19 Jan.	Big LMs: Brants *et al.* (2007); Heafield *et al.* (2013); Chelba *et al.* (2014)		
Mo, 26 Jan.	Student-suggested paper		
Mo, 2 Feb.	Student-suggested paper		

In the context of the table above, the articles having the surnames *outside* of parentheses (eg. Smith (2015)) are *primary* articles, which must be read and discussed by everyone. The articles having the surnames *within* parenthesis (eg. (Smith, 2015)) are *auxiliary* articles, which may be helpful but are not required readings.

3 Other Readings

Chen and Goodman (1998) and Goodman (2001) are great reads for when you are drifting off to sleep at night. There are other interesting statistical language models and smoothing techniques that we probably won't be able to cover in this course, unless you really want to cover it at the expense of another topic. For example:

- Trigger language models (Rosenfeld and Huang, 1992; Lau *et al.*, 1993b,a; Rosenfeld, 1994, 1996)
- Whole sentence MaxEnt models (Rosenfeld *et al.*, 2001)
- Probabilistic context-free grammars (yes, they're language models too: Baker, 1979, *inter alia)*
- Structured language models (Chelba and Jelinek, 1998; Chelba, 2000)

- Bilingual language models, for MT reordering (Mariño *et al.*, 2006; Niehues *et al.*, 2011; Garmash and Monz, 2014)

4 Links

4.1 Language Modeling Software

https://en.wikipedia.org/wiki/Language_model#External_links
(I maintain this, so hopefully it's up-to-date :-)

4.2 Free Corpora

- WMT 2014, esp. News Crawl under "Monolingual language model training data": http://www.statmt.org/wmt14/translation-task.html#download
- ACL Wiki, "Resources by Language" http://aclweb.org/aclwiki/index.php?title=List_of_resources_by_language

4.3 Corpus Processing Tools

- http://jon.dehdari.org/corpus_tools
- https://github.com/kpu/preprocess

5 Participation / Teilnahme

You will be expected to be an active part of the class. If you must miss a class, it is still your responsibility to read the materials and understand them.

6 Grading / Einstufung

- Topic presentation / Halten eines Vortrags
- Regular participation / regelmäßige Teilnahme
- Prepared questions (2x) vorbereitete Fragen
- A few small assignments, based on Koehn (2010, p. 215)
- Term paper / Hausarbeit

References

Baker, James K. 1979. Trainable grammars for speech recognition. In *Speech Communication Papers for the 97th Meeting of the Acoustical Society of America*, ed. by Dennis H. Klatt and Jared J. Wolf, 547–550, Cambridge, MA, USA.

Bellegarda, Jerome R. 2000. Exploiting Latent Semantic Information in Statistical Language Modeling. *Proceedings of the IEEE* 88.1279–1296.

Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. 2003. A Neural Probabilistic Language Model. *Journal of Machine Learning Research* 3.1137–1155.

Bilmes, Jeff A., and Katrin Kirchhoff. 2003. Factored Language Models and Generalized Parallel Backoff. In *Proceedings of the HLT-NAACL 2003 - Short Papers*, 4–6. Association for Computational Linguistics.

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research* 3.993–1022.

Brants, Thorsten, Ashok C. Popat, Peng Xu, Franz J. Och, and Jeffrey Dean. 2007. Large Language Models in Machine Translation. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 858–867, Prague, Czech Republic. Association for Computational Linguistics.

Brown, Peter F., Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-Based *n*-gram Models of Natural Language. *Computational Linguistics* 18.467–479.

Chelba, Ciprian, 2000. *Exploiting Syntactic Structure for Natural Language Modeling*. Baltimore, MD, USA: The Johns Hopkins University dissertation.

Chelba, Ciprian, and Frederick Jelinek. 1998. Exploiting Syntactic Structure for Language Modeling. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, 225–231, Montreal, Quebec, Canada. Association for Computational Linguistics.

Chelba, Ciprian, Tomáš Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, and Phillipp Koehn. 2014. One Billion Word Benchmark for Measuring Progress in Statistical Language Modeling. *Preprint* .

Chen, Stanley. 2009. Shrinking Exponential Language Models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 468–476, Boulder, CO, USA. Association for Computational Linguistics.

Chen, Stanley, and Joshua T. Goodman. 1998. An Empirical Study of Smoothing Techniques for Language Modeling. Technical Report 10-98, Harvard University.

Elman, Jeffrey L. 1990. Finding Structure in Time. *Cognitive Science* 14.179–211.

Garmash, Ekaterina, and Christof Monz. 2014. Dependency-Based Bilingual Language Models for Reordering in Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1689–1700, Doha, Qatar. Association for Computational Linguistics.

Gildea, Daniel, and Thomas Hofmann. 1999. Topic-Based Language Models Using EM. In *Proceedings of EUROSPEECH*, 2167–2170.

Goodman, Joshua T. 2001. A Bit of Progress in Language Modeling, Extended Version. Technical Report MSR-TR-2001-72, Microsoft Research.

Guthrie, David, Ben Allison, Wei Liu, Louise Guthrie, and Yorick Wilks. 2006. A closer look at skip-gram modelling. In *Proceedings of the 5th international Conference on Language Resources and Evaluation (LREC-2006)*, 1222–1225.

Heafield, Kenneth, Ivan Pouzyrevsky, Jonathan H. Clark, and Philipp Koehn. 2013. Scalable Modified Kneser-Ney Language Model Estimation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 690–696, Sofia, Bulgaria. Association for Computational Linguistics.

Huang, Xuedong, Fileno Alleva, Hsiao-Wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee, and Ronald Rosenfeld. 1993. The SPHINX-II Speech Recognition System: an Overview. *Computer Speech and Language* 2.137–148.

Iyer, Rukmini M., and Mari Ostendorf. 1999. Modeling Long Distance Dependence in Language: Topic Mixtures Versus Dynamic Cache Models. *IEEE Transactions on Speech and Audio Processing* 7.30–39.

Koehn, Philipp. 2010. *Statistical Machine Translation*. Cambridge University Press.

Kuhn, Roland, and Renato De Mori. 1990. A Cache-based Natural Language Model for Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12.570–583.

Lau, Raymond, Ronald Rosenfeld, and Salim Roukos. 1993a. Adaptive Language Modeling using the Maximum Entropy Principle. In *Proceedings of the workshop on Human Language Technology*, 108–113. Association for Computational Linguistics.

Lau, Raymond, Ronald Rosenfeld, and Salim Roukos. 1993b. Trigger-based language models: A maximum entropy approach. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP-93)*, volume 2, 45–48, Minneapolis, MN, USA. IEEE.

Mariño, José B., Rafael E. Banchs, Josep M. Crego, Adrià de Gispert, Patrik Lambert, José A. R. Fonollosa, and Marta R. Costa-jussà. 2006. *N*-gram-based Machine Translation. *Computational Linguistics* 32.527–549.

Mikolov, Tomáš, Martin Karafiát, Lukáš Burget, Jan Černocký, and Sanjeev Khudanpur. 2010. Recurrent Neural Network based Language Model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, 1045–1048.

Niehues, Jan, Teresa Herrmann, Stephan Vogel, and Alex Waibel. 2011. Wider Context by Using Bilingual Language Models in Machine Translation. In *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT)*, 198–206, Edinburgh, Scotland. Association for Computational Linguistics.

Rosenfeld, Ronald, 1994. *Adaptive Statistical Language Modeling: A Maximum Entropy Approach*. Pittsburgh, PA, USA: Carnegie Mellon University dissertation.

Rosenfeld, Ronald. 1996. A Maximum Entropy Approach to Adaptive Statistical Language Modelling. *Computer Speech and Language* 10.187–228.

Rosenfeld, Ronald, Stanley F. Chen, and Xiaojin Zhu. 2001. Whole-sentence Exponential Language Models: A Vehicle for Linguistic-statistical Integration. *Computer Speech and Language* 15.55–73.

Rosenfeld, Ronald, and Xuedong Huang. 1992. Improvements in Stochastic Language Modeling. In *Proceedings of the Workshop on Speech and Natural Language*, 107–111, San Mateo, CA, USA. Association for Computational Linguistics.

Tan, Ming, Wenli Zhou, Lei Zheng, and Shaojun Wang. 2011. A Large Scale Distributed Syntactic, Semantic and Lexical Language Model for Machine Translation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 201–210, Portland, OR, USA. Association for Computational Linguistics.