Statistical Machine Translation for Hotel Descriptions

Master Project Business Analytics

Anne Zuiker

August 29, 2015

Statistical Machine Translation for Hotel Descriptions

Anne Zuiker

August 2015

VU University Amsterdam
Faculty of Sciences
De Boelelaan 1081a
1081HV Amsterdam

Booking.com
Herengracht 597
1017CE Amsterdam

Supervisors:
Nicolaas Nobel (Booking.com)
Rene Bekker (VU University Amsterdam)
Sandjai Bhulai (VU University Amsterdam)

Preface

This thesis contains the results of using machine translation for hotel descriptions. It is the final product of a six month internship at Booking.com, which was also the Master Project of my Master programme in Business Analytics at the VU University in Amsterdam.

Hereby, I want to thank everyone who contributed to this research. First, I want to thank Nic, my supervisor from Booking.com. Thanks for believing I was suitable for the job, thanks for the guidance and for challenging every step I took. Thanks to my teammates in Content Analytics: Catherine, Dennis, Karlijn, Nimrod and Xiaowei. Your feedback and suggestions were invaluable. You have made these last six months very pleasant. Thanks to the Dutch language specialists for their help in the validation of my translations and thanks to any other colleague in Booking.com that helped me with their thoughts, ideas and feedback.

Thanks to Rene Bekker, the only one who dared to supervise me during this challenging project. You were very committed and I am very grateful for all the help with this project, that was not even in your field of expertise. Also thanks to Sandjai for taking the job of being the second reader.

On a more personal note, I want to take this opportunity to thank a few people that are very close to me. Sjanne, Hanna, Mabel, Milou and Kylie, we always told each other that we couldn't have done it without each other and I still stand by that. Thanks for making me believe in myself and for all the laughter and fun we had inside and outside the VU.

Mum & Dad, thanks for everything you have done for me. Thanks for your love, your guidance and support. You were always so involved in my study, I am eternally grateful.

Tom, thank you so much. You understood how important this all was for me. Thanks for your endless endurance and saying 'maak je niet zo druk schatje, het komt allemaal wel goed' more times than I can count.

Enjoy the read!

Bibliography

[1] About Booking.com. http://www.booking.com/content/about.en-gb.html. Accessed: 2015-08-09.

[2] De succesvolle weg die booking.com heeft uitgevonden. http://www.travelnext.nl/de-succesvolle-weg-die-booking-com-heeft-uitgevonden.html. Accessed: 2015-05-04.

[3] Hill climbing. http://artedi.ebc.uu.se/course/Embo01/Images/Summercourse14.gif. Accessed: 2015-05-15.

[4] Hybrid machine translation. https://docs.google.com/viewer?url=patentimages.storage.googleapis.com/pdfs/US20100179803.pdf. Accessed: 2015-05-01.

[5] Statistical Machine Translation and Example-based Machine Translation. http://www.proz.com/translation-articles/articles/2483/1/Statistical-Machine-Translation-and-Example-based-Machine-Translation. Accessed: 2015-05-01.

[6] Why Priceline's purchase of Booking.com was the most profitable travel deal of the 2000s. http://www.tnooz.com/article/why-pricelines-purchase-of-booking-com-was-the-most-profitable-travel-deal-of-the-2000s. Accessed: 2015-05-04.

[7] Y. Al-Onizan, J. Curin, M. Jahr, K. Knight, J. Lafferty, D. Melamed, F.J. Och, D. Purdy, N.A. Smith, and D. Yarowsky. Statistical machine translation: Final report. In *JHU Workshop 1999*, 1999.

[8] Chris Callison-Burch, Miles Osborne, and Philipp Koehn. Re-evaluation the role of bleu in machine translation research. In *EACL*, volume 6, pages 249–256, 2006.

[9] Michael Collins. Introduction to Statistical MT. https://class.coursera.org/nlangp-001/lecture/155, 2013.

[10] Michael Collins. Learning Phrases from Alignments (Part 2). https://class.coursera.org/nlangp-001/lecture/179, 2013.

[11] Michael Collins. Linear Interpolation (Part 1). https://class.coursera.org/nlangp-001/lecture/55, 2013.

[12] Michael Collins. Trigram Language Models. https://class.coursera.org/nlangp-001/lecture/69, 2013.

[13] Marta R. Costa-Jussa, Mireia Farrús, José B. Mariño, and José A.R. Fonollosa. Study and comparison of rule-based and statistical catalan-spanish machine translation systems. *Computing and Informatics*, 31(2):245–270, 2012.

[14] Marta R. Costa-Jussa and José A.R. Fonollosa. Latest trends in hybrid machine translation and its applications. *Computer Speech & Language*, 32(1):3–10, 2015.

[15] William Gale and Geoffrey Sampson. Good-turing smoothing without tears. *Journal of Quantitative Linguistics*, 2(3):217–237, 1995.

[16] William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

[17] John Hutchins. The history of machine translation in a nutshell. 2005.

[18] John Hutchins. Machine translation: A concise history. *Computer aided translation: Theory and practice*, 2007.

[19] Slava M. Katz. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *Acoustics, Speech and Signal Processing, IEEE Transactions on*, 35(3):400–401, 1987.

[20] Kevin Knight. A statistical mt tutorial workbook, 1999.

[21] Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer, 2005.

[22] Philipp Koehn. *Statistical machine translation*. Cambridge University Press, 2009.

[23] Philipp Koehn. Moses, statistical machine translation system, user manual and code guide, 2010.

[24] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[25] Philipp Koehn and Christof Monz. Manual and automatic evaluation of machine translation between european languages. In *Proceedings of the Workshop on Statistical Machine Translation*, pages 102–121. Association for Computational Linguistics, 2006.

[26] Mirjam Sepesy Maucec, Janez Brest, and Zdravko Kacic. Slovenian to english machine translation using corpora of different sizes and morpho-syntactic information. In *Language Technologies Conference: proceedings of the 9th International Multiconference Information Society IS*, volume 2006, pages 222–225. Citeseer, 2005.

[27] Tood K. Moon. The expectation-maximization algorithm. *Signal processing magazine, IEEE*, 13(6):47–60, 1996.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[29] Maja Popovic and Hermann Ney. Statistical machine translation with a small amount of bilingual training data. In *5th LREC SALTMIL Workshop on Minority Languages*, pages 25–29, 2006.

[30] Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

[31] Marco Turchi, Tijl De Bie, and Nello Cristianini. Learning performance of a machine translation system: a statistical and computational analysis. In *Proceedings of the Third Workshop on Statistical Machine Translation*, pages 35–43. Association for Computational Linguistics, 2008.

[32] Dániel Varga, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. Parallel corpora for medium density languages. *AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE SERIES 4*, 292:247, 2007.

[33] Warren Weaver. Translation. *Machine translation of languages*, 14:15–23, 1955.

[34] Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics bulletin*, pages 80–83, 1945.