Chapter 13

Learning to translate

Jaakko J. Väyrynen, Sami Virpioja, Timo Honkela, Mikko Kurimo, Marcus Dobrinkat, Tero Tapiovaara, Tommi Vatanen

13.1 Introduction

Our research on multilinguality and machine translation (MT) uses novel methods that are based on adaptivity. An MT system is *learning to translate* rather than needs to be programmed to do so. The advances in statistical machine translation have shown that the adaptive paradigm can help in reducing the system development costs dramatically. However, these systems rely on representations that do not capture many relevant linguistic aspects, neither take into account the wealth of knowledge that is known about human cognitive processes related to natural language understanding, translation and interpretation.

13.2 Analysis of complexity of European languages

We have studied differences between the European Union languages using statistical and unsupervised methods [8]. The analysis has been conducted at different levels of language including lexical, morphological and syntactic levels. Our premise is that the difficulty of the translation could be perceived as differences or similarities in different levels of language. Two approaches were selected for the analysis. A Kolmogorov complexity based approach was used to compare the language structure in syntactical and morphological levels. A morpheme-level comparison was conducted based on an automated segmentation of the languages into morpheme-like units.

13.3 Learning interlingual mappings

We have also developed an approach for finding interlingual mappings using the Self-Organizing Map (SOM) algorithm [4]. The semantic or conceptual space is explicitly modeled in the SOM-based approach. This can be constrasted with the commonly used Bayesian approach. This approach resembles, to some degree, the idea of using a knowledge-based interlingua in machine translation. The underlying philosophical assumptions about knowledge are, however, quite different. In a knowledge-based interlingua, the semantics of natural language expressions are typically represented as propositions and relations in symbolic hierarchical structures. The SOM can be used to span a continuous and multidimensional conceptual space in a data-driven manner. Moreover, the approach provides a natural means to deal with multimodal data (cf. [9]).

13.4 Applying morphology learning to statistical machine translation

Languages of rich morphology pose a problem for statistical machine translation methods, which usually apply words as the smallest units of translation. We have studied how unsupervised learning of morphology (see Section 10.2) can be used to help in the task. In a joint work with University of Cambridge [3], automatic morphological segmentations by Morfessor [1] were shown to improve the translations provided by the well-known Moses system [6]. The approach combines individual translation models that use alternative morphological decompositions using Minimum Bayes Risk decoding. Statistically significant improvents were obtained for two tasks: Arabic to English task, where two different morphological analyses were applied for Arabic, and Finnish to English, where word-based model was combined with one where Morfessor was applied for Finnish. The method was applied also in the machine translation tasks of Morpho Challenge 2009 (see Section 10.5).

13.5 Experiments in speech-to-speech machine translation

In a join effort with Speech Recognition (Ch. 8) and Natural Language Processing (Ch. 10) groups, we conducted experiments with speech-to-speech machine translation from Finnish to English [2]. The experiment is described in detail in Section 8.4.

13.6 Automatic machine translation evaluation

The feasibility of normalized compression distance as an automatic machine translation evaluation measure has been investigated [10]. The examined distance metric is based on an approximation of the Kolmogorov complexity between translated text and a reference translation. Compared to many state-of-the-art automatic measures, normalized compression distance is theoretically justified while providing competitive correlation to human judgments,.

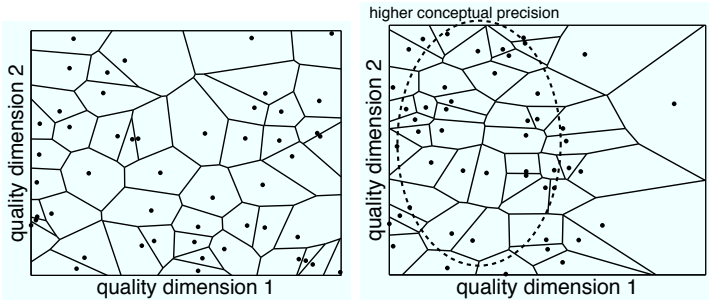

Figure 13.1: Illustration of differing conceptual densities of two agents having a 2-dimensional quality domain. Points mark the locations of the prototypes of concepts. Lines divide the concepts according to Voronoi tessellation. Both agents can discriminate an equal number of concepts, but abilities of the agent B are more focused on the left half of the quality dimension 1, whereas agent A represents the whole space with rather equal precision.

13.7 Within-language translation

The research related to machine translation includes also within-language translation activities. The basic idea is to conduct translation or paraphrasing between two different ways using the same language. In a preparatory study towards this direction, automated classification into layperson and expert use of medical language was conducted using the SVM (support vector machine) method [7].

In general, to provide motivation for this line of research, two persons may often have very different conceptual density related to a topic under consideration. For instance, in Fig.13.1 person A has a rather evenly distributed conceptual division of the space, whereas person B has a more fine-grained conceptual division on the left side of the conceptual space, but has lower precision on the right side of the space [5].

If some agents speak the *same language*, many of the symbols and the associated concepts in their vocabularies are the same. A subjective conceptual space emerges through an individual self-organization process. The input for the agents consists of perceptions of the environment, and expressions communicated by other agents. The subjectivity of the conceptual space of an individual is a matter of degree. The conceptual spaces of two individual agents may be more or less different. The convergence of conceptual spaces stem from two sources: similarities between the individual experiences (as direct perceptions of the environment) and communication situations (mutual communication or exposure to the same linguistic/cultural influences such as upbringing and education, and artifacts such as newspapers, books, etc.) [5]. In a similar manner, the divergence among conceptual spaces of agents is caused by differences in the personal experiences/perceptions and differences in the exposure to linguistic/cultural influences and artifacts. These aspects are handled in more detail in the section on socio-cognitive modeling 14.

References

[1] Mathias Creutz and Krista Lagus. Unsupervised models for morpheme segmentation and morphology learning. *ACM Transactions on Speech and Language Processing*, Volume 4, Issue 1, Article 3, January 2007.

[2] David Ellis, Mathias Creutz, Timo Honkela, and Mikko Kurimo. Speech to speech machine translation: Biblical chatter from Finnish to English. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, pages 123-130, Hyderabad, India, January 2008. Asian Federation of Natural Language Processing.

[3] Adrià de Gispert, Sami Virpioja, Mikko Kurimo, and William Byrne. Minimum bayes risk combination of translation hypotheses from alternative morphological decompositions. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 73-76, Boulder, USA, June 2009. Association for Computational Linguistics.

[4] Timo Honkela, Sami Virpioja, and Jaakko Väyrynen. Adaptive translation: Finding interlingual mappings using self-organizing maps. In *Vera KurkovÃ¡, Roman Neruda, and Jan Koutnik, editors, Proceedings of ICANN'08*, volume 5163 of *Lecture Notes in Computer Science*, pages 603-612. Springer, 2008.

[5] Timo Honkela, Nina Janasik, Krista Lagus, Tiina Lindh-Knuutila, Mika Pantzar, and Juha Raitio. Modeling communities of experts - conceptual grounding of expertise. Technical Report TKK-ICS-R24, Helsinki University of Technology, 2009.

[6] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Ondrej Bojar, Alexandra Constantin, and Evan Herb. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the ACL 2007 Demo and Poster Sessions*, pages 177–180, 2007.

[7] Marja Ollikainen. *Matching medical documents to users; Lääketieteellisten dokumenttien sovitus käyttäjille*. Master's Thesis. Helsinki University of Technology, Department of Information and Computer Science, Espoo, 2008.

[8] Markus Sadeniemi, Kimmo Kettunen, Tiina Lindh-Knuutila, and Timo Honkela. Complexity of European Union languages: A comparative approach. *Journal of Quantitative Linguistics*, 15(2):185-211, 2008.

[9] Mats Sjöberg, Jorma Laaksonen, Timo Honkela, and Matti Pöllä. Inferring semantics from textual information in multimedia retrieval. *Neurocomputing*, 71(13–15):2576–2586, 2008.

[10] Jaakko J. Väyrynen, Tero Tapiovaara, Kimmo Kettunen, and Marcus Dobrinkat. Normalized compression distance as an automatic MT evaluation metric. In *Machine Translation 25 Years on*, to appear.