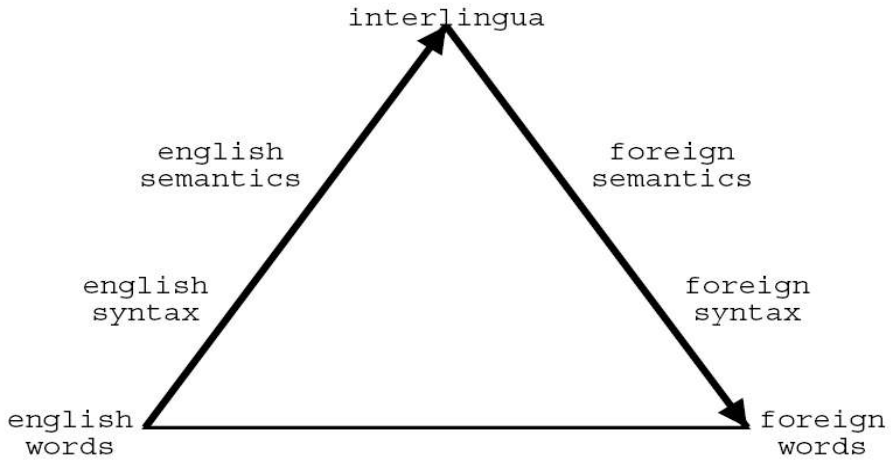
Advanced Signal Processing Seminar WS 2006

Machine Translation

Machine translation (MT) is the application of computers to the task of translating texts from one natural language to another. One of the very earliest pursuits in computer science, MT has proved to be an elusive goal, but today a number of systems are available which produce output which, if not perfect, is of sufficient quality to be useful in a number of specific domains. The following Figure shows the different types of translation models which are treated during the seminar.



In this seminar, we will try to give an overview of the state-of-the-art in machine translation, and in particular discuss current approaches and algorithms for statistical machine translation. The following list should give you an idea of how statistical MT works and which topics might be discussed in the seminar.

- Approaches to Machine Translation
  - Lexicon-based
  - Example-based
  - Interlingual
  - Statistical
- Statistical Machine Translation
  - Introduction:

    - Knight, K., and Marcu, D. *Machine translation in the year 2004*. In Intl. Conf. on Acoustics, Speech, and Signal Processing (ICASSP), 965 -968, 2005

    - Kevin Knight and Philipp Koehn, *What's New in Statistical Machine Translation*, Tutorial at HLT/NAACL 2004, 2004. (Slides are on the homepage of Kevin Knight).

  - Translation Model
    - Word-based: IBM Model 1-5: **(Sendlhofer, Wohlmayr (13.11.2006))**

- P. F. Brown, V. J. Della Pietra, S. A. Della Pietra & R. L. Mercer, *The Mathematics of Statistical Machine Translation: Parameter Estimation*, Computational Linguistics 19(2), 1993.
- Kevin Knight, "A Statistical MT Tutorial Workbook", unpublished, August 1999.

- Phrase-based: **(Reinisch, Tomescu (20.11.2007))**
  - P. Koehn, F.J. Och, and D. Marcu. *Statistical Phrase-Based Translation*. In NAACL/HLT 2003, Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference, 2003.
  - Och, F. J., & Ney, H., *The alignment template approach to statistical machine translation*. Computational Linguistics, 30, 2004.
  - D. Chiang, *A hierarchical phrase-based model for statistical machine translation*, Proceedings of ACL 2005, 2005. (best paper award)
- Syntax-based: **(Friedl, Teichmeister (4.12.2006))**
  - Yamada, K., Knight, K.: *A Syntax-Based Statistical Translation Model.* In: Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. 523-529, 2001.
  - D. Gildea. *Loosely tree-based alignment for machine translation*. In Proceedings of the 41st Annual Meeting of the Association of Computational Linguistics (ACL-03), 2003
  - Charniak, E.; Knight, K.; and Yamada, K. S*yntax-based language models for statistical machine translation.* In MT Summit IX. Intl. Assoc. for Machine Translation, 2003.

- Language Model
  (Schofield (27.11.2006))
  (Neffe, Holzmann (18.12.2006))
  - N-grams: F. Pernkopf, Speech Communication II Class
    F. Pernkopf, *Natural Language Processing with Application to Machine Translation: A gentle Introduction*, Class Notes.
  - Context-free grammar: F. Pernkopf, Speech Communication II Class
    F. Pernkopf, *Natural Language Processing with Application to Machine Translation: A gentle Introduction*, Class Notes.
  - Maximum Entropy Models:
    - A. Berger, S. Della Pietra and V. Della Pietra, *A Maximum Entropy Approach to Natural Language Processing*, Computational Linguistics, vol. 22, no. 1, 1996.
    - R. Rosenfeld, *A Maximum Entropy Approach to Adaptive Statistical Language Modeling*, Computer, Speech and Language, vol. 10, pp. 187--228, 1996.
    - Ronald Rosenfeld, Stanley F. Chen, and Xiaojin Zhu, *Whole-sentence exponential language models: a vehicle for linguisticstatistical integration,* 2001.
    - Schofield, E. , Fitting maximum entropy models on large sample spaces (http://userver.ftw.at/~ejs/schofield06fitting.pdf) and references in Section 5.1.2., PhD Thesis, 2006.

- Decoding Algorithms **(Kanzler, Pomberger (15.01.2007))**
  - Germann, U., Jahr, M., Knight, K., Marcu, D., Yamada, K., *Fast decoding and optimal decoding for machine translation.* In: Proc. of the 39th Annual Meeting of the Association for Computational Linguistics (ACL), 2001.

- Germann, U., Jahr, M., Knight, K., Marcu, D., Yamada, K., *Fast and optimal decoding for machine translation Source,* Artificial Intelligence, Vol. 154 , Issue 1-2 , 2004.
- K. Yamada and K. Knight, *A Decoder for Syntax-Based Statistical MT*,Proc. of the Conference of the Association for Computational Linguistics (ACL), 2002.

    - Optimization Techniques
        - Och, F.J., Ney, H.: *Discriminative training and maximum entropy models for statistical machine translation*. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL).,2002.
        - Shen, L., Sarkar, A., and Och, F. J. (2004). *Discriminative reranking for machine translation*. In Proc. HLTNAACL 2004.

- Evaluation **(Petrik, Sakir (6.11.2006))**
    - BLEU (BiLingual Evaluation Understudy):
      - Papineni, K.A., Roukos, S., Ward, T., Zhu, W.J.: *Bleu: a method for automatic evaluation of machine translation*. Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, 2001.
    - NIST score
      - G. Doddington, *Automatic evaluation of machine translation quality using n-gram cooccurrence statistics*. ARPA Workshop on Human Language Technology, 2002.
    - Word Error Rate (WER) (Minimum Error Rate Training):
      - Franz Josef Och. *Minimum error rate training in statistical machine translation*. In Proc. of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), 160-167, 2003.
    - Yasuhiro Akiba, Kenji Imamura, Eiichiro Sumita, Hiromi Nakaiwa, SeiichiYamamoto, Hiroshi G. OkunoUsing, *Multiple Edit Distances to Automatically Grade Outputs from Machine Translation Systems.* IEEE Transactions on Audio, Speech and Language Processing, Vol. 14, No. 2, pp. 393-402, 2006.
    - ...

- Applications **(Stadtschitz, Gutmann (22.01.2007))**
    - Overview about software, data, and tools in the internet.

Organisation

The seminar is organized by the Signal Processing and Speech Communication Laboratory of TU Graz. Students will work on a selected topic, create a self-study WWW document, and give an oral presentation in class during a 45 minute discussion session. Work in small groups of 2 or 3 students is strongly encouraged. The working language will be English.

Topics will be assigned in an introductory session on Monday, October 16, 17:00h in the Seminar Room IRT, Inffeldgasse 16c, 2nd floor (Room ID02104).