

Machine Learning in the Statistical Natural Language Processing Group of the University of Stuttgart

Alexander Fraser

Institute for Natural Language Processing
Universität Stuttgart

April 11th, 2013 - Pascal

Pascal2 at Stat NLP Stuttgart

Stat NLP group Stuttgart - led by Hinrich Schütze, sub-groups Alexander Fraser (Machine Translation), Helmut Schmid (Parsing).

Pascal2 funding played a big role in published papers on:

Pascal2 at Stat NLP Stuttgart

Stat NLP group Stuttgart - led by Hinrich Schütze, sub-groups Alexander Fraser (Machine Translation), Helmut Schmid (Parsing).

Pascal2 funding played a big role in published papers on:

- discriminative reranking of syntactic parsing (see poster)

Pascal2 at Stat NLP Stuttgart

Stat NLP group Stuttgart - led by Hinrich Schütze, sub-groups Alexander Fraser (Machine Translation), Helmut Schmid (Parsing).

Pascal2 funding played a big role in published papers on:

- discriminative reranking of syntactic parsing (see poster)
- unsupervised and semi-supervised transliteration mining

Pascal2 at Stat NLP Stuttgart

Stat NLP group Stuttgart - led by Hinrich Schütze, sub-groups Alexander Fraser (Machine Translation), Helmut Schmid (Parsing).

Pascal2 funding played a big role in published papers on:

- discriminative reranking of syntactic parsing (see poster)
- unsupervised and semi-supervised transliteration mining
- new statistical models of translation

Pascal2 at Stat NLP Stuttgart

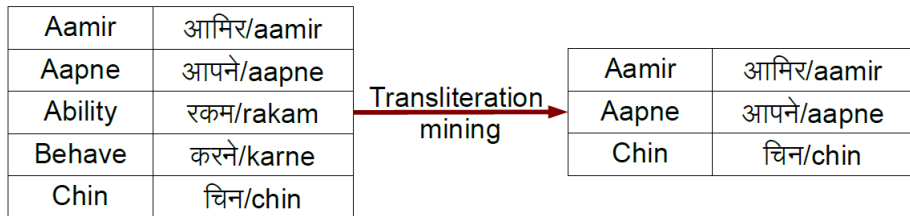
Stat NLP group Stuttgart - led by Hinrich Schütze, sub-groups Alexander Fraser (Machine Translation), Helmut Schmid (Parsing).

Pascal2 funding played a big role in published papers on:

- discriminative reranking of syntactic parsing (see poster)
- unsupervised and semi-supervised transliteration mining
- new statistical models of translation
- predicting German grammatical features when translating from English to German (see poster)

Unsupervised transliteration mining

- Consider a list of word pairs which are either transliterations of each other or non-transliterations
- This can be extracted from Wikipedia Interlanguage Links (balanced classification) or from parallel text (mostly not transliterations)
- Transliteration mining extracts the transliteration pairs from the list of word pairs
- We showed this can be done without any linguistic knowledge and without supervision



New translation models

- Long-distance reordering is a difficult problem in translation
- Google Translate (and our previous work) solves this problem using heuristic preprocessing
- We have studied long-distance reordering in two general statistical translation approaches
- Our work on **sequence-based translation**:
 - ▶ We have defined markov (n-gram) models over minimal translation units, rather than unigrams over so-called phrase pairs (phrases are memorized groups of consecutive words in two languages)
 - ▶ Our sequences contain lexically triggered word reordering which can apply over arbitrary distances
- Our work on **synchronous context-free grammars (SCFGs)**:
 - ▶ We have studied the ability of SCFGs to capture the reorderings required to translate German to English
 - ▶ We have determined how to model these reorderings in a principled way **during inference**

What is on the poster

What I didn't talk about:

- Discriminative reranking for syntactic parsing
- Predicting rich morphology (German here, but we use similar approaches when translating English to French and Spanish)

Summary:

- Overview of 5 different groups of work (within machine translation and syntactic parsing)
- Wide variety of Natural Language Processing problems, solved using machine learning techniques
- Novel generative and discriminative models, training methods, etc
- Pascal funding was very useful to us