Creating a Psychology Lexikon from a Parallell Corpus: a Pilot Study

- Outline:
  - Research Problem
  - The knowledge base
  - The study
  - Evaluation
  - Results
  - Conclusions

Research Problem

- Users value the possibility of submitting questions in their native language
- The Web4health portal supports Cross Language Question Answering with the help of Systran MT
- However Systran implements lexikons not tailored to the psychology domain
- This research tries to extract word translations from a Swedish-English corpus, i.e. a sample of Web4health, with the help of Clue Aligner tool (Tiedermann 03) in order to build a Swedish-English lexikon

The knowledge base – Web4health

- Consists of FAQs (Frequently Asked Questions) in the field of Psychology and Psychotherapy
- The content is provided by medical experts from 5 European Countries
- Every FAQ consists of question-answer pairs, where the question has a template that matches different variations of similar questions (Template-based Question Answering, Sneiders 02)

The study – The corpus selection

- The corpus of the pilot study consists of 20 FAQs in the source language (Swedish) and the target language (English)
- The Swedish corpus consists of 9089 tokens
- The English corpus consists of 8819 tokens
- The most parallel documents were chosen
- Sample corpus selection : difficult and time consuming task

The study - Research set-up

- Basic groundwork in order to prepare the texts:
  - Cleaning HTML tags
  - Conversion to XML
  - POS Tagging with TnT (Brants 00)
  - Conversion of upper cases to lower cases (after aligning at the sentence level)
- The texts in the parallel corpora were aligned at the sentence and word level
- Creation of a table with word-alignment frequencies sorted in descending order

Evaluation Method

- The results of the alignment process were assessed at the word level
- Evaluation limited at the top 800 entries in the frequency table
- The focus was on the lemmas
- In the evaluation the following fractions were calculated:
  - Fraction of *correct alignments*, e.g. ätstörningar => eating disorders
  - Fraction of *partly correct alignments*, e.g. visa upp => be show
  - Fraction of *doubles/triples*, e.g. och och => and

General Results

- Clue Aligner managed often to align successfully compound words, e.g. panikångest => panic disorder
- The system also managed to align nouns in the source language to pronouns in the target language, e.g. personer =>those
- The system had difficulties with:
  - passive forms, e.g. ”patienter uppmanas” and ”patients are asked” gave uppmanas => asked
  - genitive forms in English, e.g. individens => individual’
  - reflexive verbs in Swedish, the system did non include the reflexive pronoun in the alignment, e.g lära => learn *instead of* lära sig => learn

Summary of Quantitative Results

- Correct Alignments ≈ 71.4%
- Partly Correct Alignments ≈ 17.4 %
- Erroneous Alignments = 6%
- Alignments with Double/Triple Tokens ≈ 5.2%
- In the first 90 alignments (with frequencies between 94 and 6): only six partly correct alignments, four doubles and zero erroneous alignments
- The majority of erroneous aligments after the first 400 entries, where the frequency was two and one

From the perspective of our research…

- The system managed to identify *domain specific keywords* well, providing correct or at least partially correct alignments
- Good accuracy with domain keywords even with low frequency rate:
  - undernäring => malnutrition 1
  - ångestsymptom => anxiety symptoms 1
  - mediciner => medication drugs 1
  - panikattack => panic disorder 2

Conclusions

- The results of the study are very encouraging for the utilization of Clue Aligner as a tool to create a psychology bilingual lexikon for Cross Language Question Answering
- However the sample corpora were more ”parallel” than the complete corpora
- Clue Aligner can be used a first step in the development of the Swedish-English lexikon
- Word alignments must be combined with:
  - Manual editing
  - Stop words removal