Statistical Exploration and Exploitation of the Proximity between Swiss German Dialects and Standard German

SCHERRER, Yves

Available at:
http://archive-ouverte.unige.ch/unige:22812

Disclaimer: layout of this document may differ from the published version.
Statistical Exploration and Exploitation of the Proximity between Swiss German Dialects and Standard German

Yves Scherrer
LATL, Université de Genève
yves.scherrer@lettres.unige.ch
Outline

- The sociolinguistic and dialectal situation in German-speaking Switzerland
- The scientific background – Influences from other research work
- Available data for Swiss German dialects and for Standard German
- The problem – Automatic lexicon induction
  - Build a dictionary associating dialect words with Standard German words
- Perspectives – The relevance of the problem for Computational Linguistics and Dialectology
The Sociolinguistic Situation in German-speaking Switzerland

- **Diglossia (Ferguson 1959):**
  - Two varieties of one language (or two languages) used in functionally different contexts
    - Formal vs. informal use
    - Written vs. spoken use
  - The speakers are bilingual
  - In Switzerland, traditionally only the second distinction (communication medium) is important:
    - Dialects for spoken language use
    - Standard German for written language use
  - Recent change: Dialect use shifting to informal written language
    - 58% of E-Mails, 75% of SMS in dialect (Haas ?)
  - Dialects are not threatened at all

- **2 (conflicting) functions of dialects**
  - Intercomprehension vs. regional identification
The Dialectal Situation in German-speaking Switzerland

- **Dialectological unity**
  - All Swiss dialects belong to the Alemannic subgroup of German dialects

- **Continuum of dialects**
  - The dialects of the major cities can be viewed as prototypes.
  - In-between dialects differ more or less from these prototypes.
  - The question of the number of dialects spoken in Switzerland is inadequate.

- **Trends (Christen)**
  - Lexicon tends towards uniformisation among Swiss dialects, and between dialects and Standard German
  - Grammar (phonology, morphology, syntax, and functional words) resist to uniformisation tendencies
My Approach to Swiss German Dialects

- Because of the diglossic situation, dialects can be viewed as languages with scarce (written) resources.

- Swiss German dialects can be viewed as close cognate languages to Standard German.

- First step:
  - Focus on the relation between one dialect and Standard German.

- Additional investigations:
  - Include other dialects.
The Scientific Background

- Corpus-linguistic approach
- Machine learning
- NLP for languages with scarce resources
  - Corpus-based NLP has mainly focused on resource-rich languages like English
  - Growing need and interest in language tools for other languages (Resnik, Diab, Chiang, ... : Arabic dialects)
- Machine Translation between close languages
  - Linguistic proximity of languages facilitates machine translation – simpler methods can be used (Hajic et al.)
  - Work on recognition of cognate words in different language pairs (Mann & Yarowsky)
- Dialectometry (Nerbonne)
  - find measures to assess differences between dialects
Available Data for Swiss German Dialects

- **Text corpora:**
  - Dialect literature
    - mostly Bern dialect (tradition)
    - dialectal variation
    - orthographic variation (but not inside one text)
  - Transcripts of oral interactions
    - interviews made for dialectological research purposes
    - standardized, phonologically transparent transcriptions
    - usual "noise" of spontaneous speech
      - repetitions, hesitations, pragmatically shortened utterances
  - Blogs, Internet pages, Wikipedia
    - important dialectal and orthographic variation
    - variable quality
Available Data for Swiss German Dialects

- Dictionaries, word lists:
  - Dictionaries for different dialects:
    - mostly created in the early 1980s, not digitalized
    - created from a dialectological point of view:
      - limited to lexical words
      - limited to words which differ heavily from their Standard German equivalents (non-cognate words)
      - no inflected word forms
  - Word lists available on the Internet:
    - quite small
    - different dialects
    - variable quality
    - same limitations as above
Available Data for Standard German

- High quality symbolic parsers
- High quality monolingual lexicon at LATL
- Corpora for training statistical NLP tools
- Tree Banks (syntactically annotated corpora)
  - TIGER
  - NEGRA
- Corpora with semantic role annotation
  - SALSA
Available Data

Some Conclusions

- Massive variation in Swiss German data needs a robust system
  - Hypothesis: A statistical approach will perform better than a rule-based, discrete approach.
  - To work with usable data for syntactic analysis, I choose literary texts and graphematic (vs. phonetic) representation of words.
  - Depending on results, I may use phonetic representations in a second step.

- The diglossia situation implies that (almost) no parallel corpora (Standard German + dialects) are available.
  - Parallel corpora are typically created by translation. Because speakers are bilingual, translation is limited to special situations, thus rare.
  - Hypothesis: The similarity between dialects and Standard German can leverage this disadvantage.
Data are much more readily available for Standard German.

Hypothesis: The Standard German data can be used to improve the learned dialect data.
The Problem
Automatic Lexicon Induction

- For a given dialect word, what is the most probable Standard German translation?
  - süsch > sonst  \textit{(Bern dialect examples)}
  - Prüefig > Prüfung
  - zie > ziehen
  - uschoche > auskochen
  - bblybe > geblieben

- Task:
  1. Take a dialect word and generate a lot of slightly modified translation candidates (words and non-words).
  2. Filter these candidates with a monolingual Standard German lexicon. Keep only German words.
  3. Filter the remaining candidates according to their syntactic (and semantic) role. Keep only the words that "fit into the sentence". (That's why we need syntactically analyzable raw data.)
Automatic Lexicon Induction
An Example

- Example:
  - *I ha hunger.* "I am hungry."
  - *Ich habe Hunger.*

1. Generate slightly modified translation candidates for *ha*:
   - *he, *hk, *ho, im, nah, hab, hat, *cha, gabe, habe, ...

2. Eliminate non-words:
   - *im, nah, hat, gabe, habe, ...

3. Eliminate syntactically and semantically incompatible words:
   - *habe*

- Ideally, get one result.
Automatic Lexicon Induction
First step

- How to construct the translation candidates?
- What are "slightly modified" words?
  - good translation:       bad translation:
    - speter > später     vs.     speter > *spkter
    - mache > machen   vs.     mache > *macheg
    - Sägu > Segel     vs.     Sägu > Säge
    - Hut > Haut       vs.     Hut > Hut

- Rules of phonetic changes must be respected
- Three models:
  - no phonetic preferences
  - hand-made rules (transducer)
  - automatically learned rules (stochastic transducer)
Automatic Lexicon Induction
First Step

- String edit distance (Levenshtein):
  - Substitute 1 character by another one: cost 1
  - Add 1 character: cost 1
  - Delete 1 character: cost 1

- Improvements (Nerbonne):
  - V > C is more costly than V > V
  - Include phonological context (bigrams, trigrams)

- Automatically learned rules (Ristad & Yanilos; Mann & Yarowsky):
  - Not all costs are equal. Learn different edit costs for:
    - the 3 operations
    - the letters (phonemes) involved
    - the context involved
  - Give the computer a list of word pairs to determine the frequency/probability/cost of the operations (training).
Building a lexicon is a preliminary task for the construction of a machine translation system.

Mann & Yarowsky: Such a lexicon is sufficient to obtain a satisfactory machine translation system.

This lexicon can be used as alternative or complement to existing dictionaries:

Alternative: Faster than scanning an existing typeset dictionary.

Complement: Use an existing dictionary for non-cognate words.

This lexicon can be used to facilitate the word-by-word alignment of parallel corpora.

Due to the proximity of the languages, automatic construction of a (low-quality) parallel corpus is possible, which can then be used for syntactic and semantic annotation projection (Lapata & Padó).
Automatic Lexicon Induction
The Importance of the Method

- The transducers represent models of phonological change rules.
  - They are part of a model of a bilingual speaker.
- In fact, Swiss Germans frequently use such a model in both directions:
  - Dialect speakers adapt Standard German words (technical terms, neologisms) to dialect pronunciation and morphology.
  - Children learning Standard German build Standard German words from known dialect words (overgeneralizations are observable).
Automatic Lexicon Induction
The Importance of the Method

- The transducers represent the distance between a dialect and Standard German.
  - Different transducers represent distances to different dialects.
  - We can compare different dialects by comparing different transducers.

```
  Std German
   /   \
  /     \
Bern dialect  Zürich dialect
  \
  Std German
   /     \
   /      \
Zürich dialect

  Std German
    / \
    /  \
  * Common Swiss German
    /     \
   /       \
Bern dialect  Zürich dialect
```