Preparation of standardized data from conventional dictionaries and corpora

Alla Münch

Lexicon Formalisms

30.06.2016
Overview

1 Learning from conventional sources
2 Learning from conventional dictionaries
3 Learning from corpora
4 Combining corpus and dictionary data
5 Conclusions
Table of Contents

1 Learning from conventional sources
2 Learning from conventional dictionaries
3 Learning from corpora
4 Combining corpus and dictionary data
5 Conclusions
Lexical resources:

1. for human users
   - print dictionaries
   - electronic dictionaries
   - corpora

2. for computer programs
   - demand for large lexicons for NLP purposes
   - syntactic, semantic, pragmatic, morphological and phonological information is to be included

How can information from conventional (human-oriented) resources be used as a basis for computer-oriented lexicons?
Learning from conventional sources

Three cases studies:


■ Didakowski & Geyken (2013) “From DWDS corpora to a German Word Profile – methodological problems and solutions”

Table of Contents

1 Learning from conventional sources

2 Learning from conventional dictionaries

3 Learning from corpora

4 Combining corpus and dictionary data

5 Conclusions
The dictionary of German idioms (H. Schemann 1993):

- 32,000 entries
- idioms in a very broad sense (from completely frozen expressions to multi-word expressions where all components are substitutable)
- an entry is not a linear multi-word expression but rather a set of different syntagmatic contexts around the target word
- the complexity of form varies from two-word expressions to complex PP-VP expressions

→ How to extract information from this dictionary (e.g., POS sequences) for NLP purposes?
Abschluß: etw. zum/zu einem Abschluß bringen *form*
1. Wenn du die Arbeit jetzt nicht zum Abschluß bringst, wird sie nie mehr fertig. Denn im nächsten Jahr hast du dafür keine Zeit mehr. Also, halte dich jetzt dabei, daß du in aller Kürze fertig bist.
2. Was man anfängt, sollte man auch zum Abschluß bringen.

zum Abschluß kommen/gelangen *form*
Sind die Verhandlungen endlich zum Abschluß gekommen? – Sie wurden gestern beendet.

abschminken: das/(die Hoffnung/…) kann ich mir/kann der Peter sich/… abschminken ugs
Die Karin und dir helfen!? Daß kannst du dir abschminken! Du hilfst auch niemandem.

abschneiden: gut/schlecht/günstig/… abschneiden (bei einem Examen/…)

abschreiben: jn./etw. abschreiben können ugs
1. Jetzt fehlt der Klaus doch schon wieder! Will er bei den Proben zu unserem Theaterstück nicht mehr mitmachen? – Ich glaube, den können wir abschreiben. Vor ein paar Tagen hat er dem Dieter erlegnet einen wahren Hammer abgekauft. Wo hat der dann hinterher denn seine Augen, wenn er so was nicht sieht?

jn. ins Abseits drängen ugs Neol
Die Mehrheit der Partei versucht ganz offensichtlich, den Baumann ins Abseits zu drängen. Er ist ihnen zu unabhängig: also darf er keinen Einfluß haben …

ins Abseits geraten ugs Neol

im Abseits stehen Fußball u. ä. – (chcr.) abseits sein/stehen

Absicht: etw. mit (voller) Absicht tun
… wenn es aus versehen geschehen wäre! Aber ich habe genau gesehen, daß er sie mit voller Absicht ins Gesicht gestoßen hat.

etw. in guter/bester/böser/… Absicht tun
Ich gebe zu, sie hat dir mit ihrer Bemerkung geschadet. Aber sie hat das doch nicht in böser Absicht getan …

nicht die Absicht haben, sich etw. bieten zu lassen/…
… Ich habe nicht die Absicht, mich von Ihnen beleidigen zu lassen! Entweder Sie ändern Ihren Ton oder Sie gehen!

die erste Absicht haben, etw. zu tun
Der Herbert hat die erste Absicht auszuwandern. Er sagt, von diesem Ende sei alles jetzt ab, wie ab.
Problem 1: multi-word entries of a paper dictionary cannot be extracted as a simple string

Examples of entries:

1. das/ etw. ist ein dicker Hund
   *that/ sth. is a fat dog*
   “that’s outrageous”
   \[→\] das ist ein dicker Hund

2. Von A bis Z Unsinn / erlogen / erfunden ... sein
   *From A to Z nonsense / lied / invented ... be*
   “Sth. is a pack of lies from A to Z”
   \[→\] von A bis Z Unsinn sein
   \[→\] von A bis Z erlogen sein
   \[→\] von A bis Z erfunden sein
Problem 2: dictionary entries cannot be parsed in the same way as naturally occurring text

Proposed method:

- Machine Learning approach: encode a certain number of entries by hand, qualify them as a training corpus, and then train an appropriate tagger on these rules in order to obtain appropriate parameter files
- a very simple tagset consisting of 10 tags
- a training set of approximately 6,000 dictionary entries tagged by hand
- the training set contained 230 different POS sequences (construction classes)
Association of a dictionary entry with a construction class:

- there exists an isomorphism between the construction class of the training set and the computed (elementary) POS sequence, or

- a complex entry can be decomposed into several elementary entries and more than 50% of the computed POS sequences are isomorphic to a single elementary construction class.

Von A bis Z Unsinn / erlogen / erfunden ... sein:

von A bis Z Unsinn sein
→ prep N prep N N V

von A bis Z erlogen sein
→ prep N prep N V V

von A bis Z erfunden sein
→ prep N prep N V V
Association failure example:

das/(die Hoffnung) kann ich mir/ kann Peter sich/... abschminken
that/(the hope) can I me/ can Peter himself/... forget about
“l/Peter/... can forget the idea/my/his/... hopes/... (of doing
sth.)”

Output:
das kann ich mir abschminken
→ det V Pron Pron V
das kann ich mir Peter sich abschminken
→ det V Pron Pron N Pron V
die Hoffnung ich mir abschminken
→ det N V Pron Pron V
die Hoffnung kann ich mir Peter sich abschminken
→ det V Pron Pron N Pron V
55% of the Schemann entries were successfully assigned to POS-sequence models.

Reasons for association failures:
- The entry is too complex (ambiguous scope of the slash (/))
- The computed POS sequences do not correspond to any of the construction classes identified in the training set
- Transcription errors due to typewriting or incorrect segmentation of the entries

Solutions:
- Manual disambiguation of the slash
- a priori definition of plausible construction classes

→ recall rose to 86% (by more than 30%)
Geyken & Boyd-Graber (2004): Learning from a Dictionary of Idioms

Conclusions:
- a relatively small training set is sufficient
- no language-specific linguistic resources required except for the training corpus
- considerable gain in time compared to manual tagging

Applications:
✓ a good basis for a program that recognizes idioms in unrestricted texts
✓ could be extended to other print dictionaries containing multi-word expressions
  ■ the major adaptation: description of the entry format of different dictionaries
  ■ useful for print dictionaries in languages that have historically been underdeveloped with respect to NLP
Wortprofil 2012:

- a lexical profiling tool for German based on grammatical co-occurrences
- implemented as an additional functionality of the lexical information system of the Digital Dictionary of the German Language (DWDS)
- provides separated co-occurrence lists for twelve different grammatical relations and links them to their corpus contexts
- the co-occurrence lists and their ordering are based on statistical computations over a fully-automatic annotated German corpus
Didakowski & Geyken (2013): Learning from corpora

Applications:
- assists lexicographers in compiling collocations
- provides useful corpus-based syntagmatic information for users interested in improving their language production skills

Examples of questions:
- ✓ Which attributive adjectives are typically used for the noun “Vorschlag” (‘proposal’)?
- ✓ Which active subject does a verb like “ausstoßen” (‘emanate’) usually take?
Didakowski & Geyken (2013): Learning from corpora

Corpus:

- German daily and weekly newspapers
- 1,783,530,826 tokens
- from 1900 up to now
- the texts are automatically tokenized, split up into sentences and annotated with part-of-speech tags

Syntactic relations:
prepositional group, adjective attribute, modifying adverbial, active subject, accusative object, coordination, genitive attribute, verb prefix, predicative complement, dative object, passive subject, comparative conjunction
Challenges of extraction of German grammatical co-occurrences:
- variable placement of the finite verb (verb-first, verb-second and verb-final)
- relatively rich morphology
- a relatively flexible phrase ordering (discontinuous verb chains, separable prefixes)
- very high ambiguity rate with regard to the morphological case information

Cf.: Sketch Engine (Kilgarriff and Tugwell, 2002) uses patterns over part-of-speech sequences to detect grammatical relations (does not work well for languages with free word order and rich morphology)
Relation extraction:
syntactic parsing as backbone for the extraction of grammatical co-occurrences:

- the sentences of a corpus are syntactically analysed via a robust dependency parser
- the syntactic relations are extracted from the parsing results

Example:
‘Der Mann nimmt den Hörer ab’
(‘The man picks up the earphone.’)
→ three binary grammatical relations relevant for collocation extraction:

1. verb prefix: <nehmen ‘to pick’, ab ‘up’>
2. active subject: <abnehmen ‘to pick up’, Mann ‘man’>
3. accusative object: <abnehmen ‘to pick up’, Hörer ‘earphone’>
Didakowski & Geyken (2013): Learning from corpora

Parser details:
- Rule based dependency parser SynCoP (Syntactic Constraint Parser, (Didakowski, 2008))
- only assigns dependency structures which are allowed by its grammar
- it is possible to assign weightings to the grammar rules
- allows partial analyses in order to annotate as much as possible
Statistical computations:

- logDice based on the Dice coefficient (Rychlý, 2008)
  \[
  logDice = 14 + \log_2 \frac{2f_{xy}}{f_x + f_y}
  \]

- MI-log-Freq based on mutual information (Kilgarriff and Tugwell, 2002)
  \[
  MI \log \text{Freq} = \log_2 \frac{f_{xy}N}{f_xf_y} \cdot \log f_{xy}
  \]

- \( f_x = \text{number of occurrences of word X} \)
- \( f_y = \text{number of occurrences of word Y} \)
- \( f_{xy} = \text{number of co-occurrences of words X and Y} \)
- \( N = \text{size of the corpus} \)
Didakowski & Geyken (2013): Learning from corpora

Generation of a sorted candidate set of collocations (ordering by collocational strength):
→ a ranking approach and a threshold approach are combined:
  ■ threshold 5 for absolute frequency of word occurrence
  ■ the candidate set is sorted by association score
Sample output:

Didakowski & Geyken (2013): Learning from corpora
Didakowski & Geyken (2013)
Learning from corpora

- extraction of statistically salient and syntactically relevant co-occurrences from a corpus
- extraction method is based on syntactic and statistical analysis
1 Learning from conventional sources
2 Learning from conventional dictionaries
3 Learning from corpora
4 Combining corpus and dictionary data
5 Conclusions
Aim: to combine the structured but incomplete information from Machine-Readable Dictionaries (MRDs) with the unstructured but more complete information available in corpora for the creation of a bilingual database

→ automatic lexicon construction integrating information from different sources into one data structure
BICORD system (Bilingual CORpus-enhanced Dictionaries) linking entries in Collins EN-FR and FR-EN bilingual dictionary with a large bilingual corpus (Canadian Parliamentary Proceedings)

BICORD can be used in two directions:
- to enhance a MRD with statistical data
- to enhance a statistical system with data from a MRD
Case Study: motion verbs

Challenge:
Manner of movement is part of semantics of English verbs, in French it is a separate surface unit:

amble – marcher d’un pas tranquille
whiz – aller a toute vitesse en sifflant

(problem for establishing lexical correspondences, as the correspondence is not one-to-one)
Initial list of verbs:
Motion verbs are extracted from the Webster taxonym dictionary based on their hypernyms:

drift – to move or float smoothly and effortlessly
walk – to move along on foot
limp – to walk lamely
...
→ initial list of 500 verbs

Problem: polysemy (cf. “run” as hypernym)
→ only verbs with 2-25 dictionary senses were taken
Search through the bilingual dictionary:

- Selected items were used to search the EN-FR dictionary for translations
- The FR-EN dictionary was searched for French headwords with the English word in the translation field (to expand the list of translations)
Extraction of relevant citations from the corpus:

- search based on the obtained list
- a simple morphological system used to expand baseforms into inflected forms to query the corpus
- POS-tagging was applied to the English side of the corpus to cull out verbal usages (not relevant for French)
  Cf. amble – marcher d’un pas tranquille

→ the set of English citations containing the verb in any morphological form and a corresponding French sentence
Creating lexical data:

- extraction of citations that have lexical correspondences already provided by MRD from the corpus
- construction of an extended lexicon using the structure provided by the bilingual dictionary with computed frequencies over matches
- possibly adding information based on observed co-occurrences

Issues:

- structure of the lexical entry
- using sublanguage corpora
- significance threshold for the correspondences
Applications:

- for lexicographers
- for the users of dictionaries
- for translators
Conclusions

- What can be learned from conventional sources of lexical information?

- Where can the information be found?

- How can it be learned?
Conclusions

- What can be learned from conventional sources of lexical information?
  - recurrent patterns
  - statistical distributions
  - collocations
  - inter-language correspondences
  - ...

- Where can the information be found?

- How can it be learned?
Conclusions

- What can be learned from conventional sources of lexical information?
  - recurrent patterns
  - statistical distributions
  - collocations
  - inter-language correspondences
  - ...

- Where can the information be found?
  - corpora
  - bilingual dictionaries
  - monolingual dictionaries
  - ...

- How can it be learned?
Conclusions

■ What can be learned from conventional sources of lexical information?
  ✓ recurrent patterns
  ✓ statistical distributions
  ✓ collocations
  ✓ inter-language correspondences
  ✓ ...

■ Where can the information be found?
  ✓ corpora
  ✓ bilingual dictionaries
  ✓ monolingual dictionaries
  ✓ ...

■ How can it be learned?
  ✓ machine learning
  ✓ parsing
  ✓ ...


