Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

Lexical Acquisition

Lothar Lemnitzer, Claudia Kunze

lothar@sfs.uni-tuebingen.de, kunze@sfs.uni-tuebingen.de

Computational Lexicography at ESSLLI 2005

Topics

- Definition
- Motivation
- Preliminaries
- Tasks
- Case study

Definition

Lexical Acquisition deals with how to obtain, with computational methods, information about the lexical units of a language from texts in this language.

Motivation

- High quality and broad coverage lexical information is needed for many NLP applications.
- Traditional lexical resources are often not available for machine processing
- If available, they did not fulfill the needs and expectations of the community
- Very large, broad coverage corpora are available or in reach nowadays
- The focus is therefore on methods to acquire the information from corpora

Shortcomings of MR dictionares

As machine-readable dictionaries became available for NLP, the information acquired from them turned out to be:

- too old
- too narrow in scope (e.g. no proper names)
- inconsistent (both within one dictionary and across dictionaries)
- missing important information items (e.g. frequency and distribution information)
- biased towards infrequent phenomena (listing infrequent phenomena, including obsolete senses)
- unreliable

Conclusion

Seminar für Sprachwissenschaft
EBERHARD KARLS UNIVERSITÄT TÜBINGEN

It turned out to be better to invest efforts in lexical acquisition than to exploit machine readable dictionaries.

Preliminaries

In order to successfully identify and extract lexical information from text corpora

- one has to define clear-cut linguistic or lexical categories
- one has to define the type of (lexical) sign which should be investigated

Lexical categories

- LCs are used to classify lexical elements on the basis of some common features
- Membership in a lexical category determines the linguistic features of the lexical item (e.g. distribution, possible functions)
- a system of LC should be defined a priori, not as a result of corpus studies

Lexical types

- The lexical type to be searched determines the complexity of corpus analysis
- Construction types: word part (morpheme), word, multi-word unit
- Grouping: single word form or word form paradigm of a lexical unit
- Some choices imply preprocessing of the corpus or more sophisticated search tools

Types of lexical acquisition - Identification

- One acquisition task is to identify or verify the existence of lexical items
- Applications: lexicography, named entity recognition
- The task is easy to solve in general; the identification of idioms requires some sophistication

Types of lexical acquisition - Classification

- Once identified, lexical items must be classified, according to some criteria
- Distinctive features for building classes are needed; they should be established on linguistic grounds; ideally, classes can be distinguished in corpora (by their distribution, their cotexts etc.)
- Applications: (semi-)automatic extension of lexical resources, corpus annotation

Types of lexical acquisition - Relations

- Single lexical units enter lexical-semantic relations
- Lexical level: selectional preferences; Textual level: lexical chains
- Applications: lexical semantics, information retrieval

A method for lexical acquisition projects

- definition of the task (which kinds of items / features should be extracted)
- selection of the data sources (which corpora; how large; which register)
- decision about the lexical categories and lexical types to be investigated
- definition of an extraction method (clustering, (un)supervised learning etc.)
- implementation of the extraction method
- data analysis
- evaluation and improvement of the method

A case study

We chose the work of Merlo and Stevenson (2001) on automatic verb classification as an example of good practice

- The linguistic categories to be analysed are chosen a priori
- For known elements of these classes, salient features are explored
- These salient features are used to classify new lexical items (learning)

Selection of data sources

The complexity of the task – the features to be identified – calls for annotated corpora

- 65 million word corpus, automatically tagged
- 29 million word subset, automatically parsed
- The extraction methods have to deal with annotation errors

Linguistic categories

The authors want to distinguish three optionally intransitive verb classes:

- Unergative verbs (*The horse raced past the barn*)
- Unaccusative verbs (*The butter melted in the sun*)
- Object-Drop verbs (*The girl played*)
- All verbs have also transitive uses
- Note that there are no differences in the surface distribution of the verbs
- Motivation for these classes: linguistic, machine translation, text generation

Lexical items

60 verbs the classes of which is known are chosen for the definition of discriminating features:

- Unergative class (semantic class: MANNER-OF-MOTION, *jumped, rushed, marched*)
- Unaccusative class (semantic class: CHANGE-OF-STATE, *opened, exploded, collapsed*)
- Object-Drop class (several semantic classes, *played, kicked, inherited*)

Extraction method

- Take into account both transitive and intransitive uses of the verbs
- Search for quantifiable, discriminating features for the classes
- Apply these features to classify new verbs
- The method is an instance of supervised learning

Relevant features

- markedness of transitive structure, correlates with frequency (Unerg < Unacc < ObjDrop)
- Frequency of causative use (Unerg, ObjDrop < Unacc) $\rightarrow$ occ. of same NP as subject and object of the verb
- Animacy of subject (Unacc (Theme subject) < Unerg, ObjDrop)
- Use in passive voice (Unerg < Unacc < ObjDrop)
- Verb used as participle (Unerg < Unacc < ObjDrop)

Implementation

The main task is identification and counting of structural patterns in the corpora (T = tagged corpus; P = parsed corpus)

- TRANS: some word classes following the verb were taken as indicators for an object (T)
- PASS: a ’VBN’ (past participle) verb in neighborhood to a form of *to be* was taken as passive (T)
- VBN: the ’VBN’/’VBD’ (past tense) ratio for the verb was calculated (T)
- CAUS: extraction of subjects and objects for a verb and measuring the overlap (P)
- ANIM: animacy was only assigned to pronouns. Calculation of ratio pronouns/NPs in subject

Implementation

- verbs are represented as vectors containing the verb name, the quantitative values and the verb class
- verb data have been divided into training and test set (single-hold-out method)
- classification is based on vector comparison
- the discriminative power of single features and features bundles have been tested (better than each feature individually)
- all features and all feature without PASS ('passive') performed best

Evaluation

- Best results were 70 % of correct class assignment
- Lower bound: 33 %
- Upper bound; 87 % (classification by human experts)

Further findings

- the features are not optimal for the discrimination of UNERG and OBJDROP verbs
- highest-frequency verbs are the most difficult cases
- improved annotation boosts precision of the classification

Problems

- to classify an unknown verb, one needs much data which will most probably not be available for neologisms
- the proposed method classifies verb lexemes and thus does not capture (regular) polysemy