

Data-Driven Methods for Acquiring Linguistic Information

Aline Villavicencio and Valia Kordoni
(UFRGS, Brazil and UdS, Germany)

in collaboration with

Timothy Baldwin and Anna Korhonen
(University of Melbourne, Australia and
Cambridge University, UK)

Motivation

- Combination of deep symbolic analyses and machine learning techniques for adequate performance in applications as Machine Translation (MT), Question-Answering (QA), Information Extraction (IE), Information Retrieval (IR), etc
- Limitation in coverage of linguistic resources
- Typical sources of coverage deficiency
 - unknown words
 - words for which no syntactic or semantic category in the dictionary
 - missing grammatical constructions

Motivation (cont.)

- The manual extension of resources is
 - costly
 - time consuming
 - error-prone

ESSLI 2006 course

- The 18th European Summer School in Logic, Language and Information
<http://essli2006.lcc.uma.es/>
- Data-Driven Methods for Acquiring Linguistic Information
<http://essli2006.lcc.uma.es/give-page.php?id=5>
- The course aims to imbue participants with an appreciation of
 - the challenges faced by data-driven approaches to linguistic knowledge acquisition, and
 - the state-of-the-art methods and tools which tackle these issues

An Introduction to Computational Word Learning

- What is Computational Word Learning?
Working definition: the process of discovering word commonalities, embellishing information contained in existing lexical resources and/or expanding existing lexical resources through computational means
- Computational word learning is also known as lexical acquisition

Brief History of Word Learning

- Word learning first came to prominence in the late 1980's as electronic lexical resources (LRs) became available
- Early efforts focused on extracting “knowledge” out of LRs
- Shift in the late-1990's towards corpus-based approaches

Recurring Themes in Word Learning

- Corpora
- Supervised vs. unsupervised methods
- Tokens and types
- Ambiguity and disambiguation
- Words and multiword expressions (MWEs)

Corpora

- A corpus (plural corpora) is a body of written or spoken language, generally either from a homogeneous source or balanced across multiple sources in an attempt to be representative of a given language type
- Examples: British National Corpus (BNC), Penn Treebank (Brown, WSJ, Switchboard)

Supervision

- Supervised methods have prior knowledge of a closed set of word classes and set out to discover and categorise new words according to those classes
- Unsupervised methods dynamically discover the word classes in the process of categorising the words
- Supervised or unsupervised? James Cook, George de Mestral (inventor of Velcro), David Livingstone

Types and Tokens

- The no. of types in a corpus is the no. of unique word forms, and the no. of tokens is the total word count

Pease porridge hot

Pease porridge cold

Pease porridge in the pot

Nine days old

- Types: 10 (*Pease, porridge, hot, cold, ...*)
- Tokens: 14 (*Pease, porridge, hot, Pease, ...*)

Ambiguity and Disambiguation

- Ambiguity: observation that a given word occurs in multiple configurations
- Disambiguation: determination of which of a fixed set of classes a given word conforms to

The gang held up the bank

The boat pulled up at the bank

We stopped by the bank

Words and Multiword Expressions

- (*Escapist definition*) A word is what we would expect to occur as an atomic, independent entry in a dictionary (e.g. *reconsider*)
- (*Narrow definition*) A multiword expression (MWE) is made up of multiple words and is syntactically and/or semantically idiosyncractic (e.g. *look up, phone book, off screen*)

Basic Topics

- Word discovery
- Morphology
- Subcategorisation Frames
- Selectional Preferences
- Diathesis Alternations
- Lexico-syntactic clues to similarity
- Semantic compositionality
- Noun countability
- Compound nominals
- Conceptual Ambiguity and Disambiguation

Word and MWE Discovery

- Segmentation (“word splitting”) of non-segmenting languages (e.g. *spotthebreaksinthisstring*)
- Extraction of collocations and multiword expressions (MWEs) (e.g. *pick out the MWE*)
- Morpheme discovery (e.g. *antidisestablishmentarianism*)

Morphology

- When are words with the same stem also related semantically?
 - Organization vs organ
 - University vs universe

Subcat Frame Acquisition

- What complements does a given word (verb/noun/adjective) take?

give [Kim]NP [a present] NP

refer [to the article] PP

fondness [for chocolate] PP

- How to distinguish between arguments and modifiers computationally

Selectional Preferences

- What set of things can normally *fly*?
? *Eg. birds, aircraft, missiles*
- How do we learn sets like this from corpora?

Diathesis Alternations

- Diathesis alternation: systematic valence-level correspondence between subcat frames of a given verb (or verb paradigm)
Kim opened the door - The door opened
- Corpus- and dictionary-driven methods for extracting verbs which participate in given alternations

Diathesis Alternations

- Diathesis alternations in verb clustering
 - Give vs donate
 - Kim gave the computer to the school
 - Kim donated the computer to the school
 - AND
 - Kim gave the school the computer
 - BUT NOT
 - Kim donated the school the computer

Noun Countability

- Noun countability: lexical property that determines which uses a noun can occur in

a/one dog, two/many/some dogs

∅/much/some information

- Corpus-based methods for determining the countabilities of English nouns
- Cross-linguistic countability prediction: English-Dutch

Lexico-syntactic patterns

- Lexico-syntactic pattern: a recurring pattern which can often signify a regular semantic relationship between some of its constituents

France and Spain and France it and other European countries

- How reliable are such patterns?
- How can the relationships be checked / validated?

Semantic Compositionality

- The degree of semantic compositionality/idiomaticity varies across MWEs:

red army, red dwarf, red herring

chicken out, bow out, make out

- Different methods for estimating the relative
- compositionality of MWEs

Conceptual Ambiguity and Disambiguation

- Nearly all the words we might learn are potentially ambiguous
- Even when we don't think of them as ambiguous
- How do we learn
 - which words are ambiguous?
 - when a particular meaning is used?
 - when does it matter?

Compound Nominals

- Compound nominals are highly productive and semantically varied:
diesel truck/tanker, phone book, apple juice seat, cloud bus
- How is it possible to constrain the range of interpretations and determine the default interpretation for a nominal compound in isolation and in context?

Automated Lexical Acquisition

- van Noord (2004)
- Baldwin (2005)
- Zhang and Kordoni (2005)
- van Noord (2006)

Error Mining and Automated Lexical Acquisition (van Noord 2004, 2006)

- An algorithm to automatically extend the lexicon
- Errors are found by error mining (van Noord 2004)
- Sentences with missing or incomplete lexicon entry are parsed with a 'universal tagset' for the unknown word
- Maximum entropy classifier is used to extract the correct tags

Automated Lexical Acquisition with secondary LRs (Baldwin 2005)

- Based on a set of lexical types
- Treat lexical acquisition as a classification task
- Generalize the acquisition model over various secondary language resources

POS tagger

Chunker

Treebank

Dependency parser

Lexical ontology

Automated Lexical Acquisition for Open Text Processing (Zhang&Kordoni 2005)

- Error mining based lexical error detection

Experiment with ERG and BNC shows that a major part of parsing failure is due to missing lexical entries.

Some rules are used to discover missing lexical entries.

- Statistical lexical acquisition

A maximum entropy based lexical type prediction model is designed and evaluated with various feature templates for various grammars.

Disambiguation model is incorporated to enhance robustness.

Learning Missing Grammatical Constructions

- Inductive Logic Programming (ILP) = Cussens and Pulman (2000) have shown that it is possible to learn complex unification grammar rules using this method, embodied within a chart parsing framework.
- One can think of a bottom up chart parser as an inferential system deriving a 'sentence' theorem from a sequence of word 'axioms' and grammatical 'inference' rules.
- Casting this within ILP, we can reason backwards from a failed parse (on the assumption that the input was grammatical) to isolate the missing rules or lexical entries needed to complete the parse.

Learning Missing Grammatical Constructions (2)

- What is needed is a rich ‘background’ theory with which to filter these hypotheses.
- One possibility that can be explored is using classifiers of parse success in written and spoken language to classify the re-parsings of the failed sentences with each of the different hypotheses.
- A classifier trained on correct parses ought to be able to detect whether a parse resulting from a newly hypothesised rule is a ‘good’ one or not.
- Other measures that one can experiment with might be to develop some kind of trainable ‘distance measure’ between hypothesised and attested rules.
- Rules that might equally well produce parses for the same data could be distinguishable along this dimension.

THANK YOU FOR YOUR ATTENTION!