

What Can Be Learnt and Acquired from Non-disambiguated Corpora

G. Nenadic, I. Spasic, S. Ananiadou

University of Salford, UK
University of Belgrade, YU

<http://www.cs.salford.ac.uk/NLP.htm>

Outline

- Introduction
learning from non-disambiguated corpora
- Case studies
 - lexical acquisition
 - learning grammatical constraints
 - semantic acquisition
- Conclusions and further research

Non-disambiguated corpora

- Larger corpora are usually not disambiguated
 - corpora for minority languages
 - domain-specific corpora
- Main source of ambiguity is homography
 - inflectional homographs
e.g. 50% of words in text are ambiguous in Serbian
gospodja: N:fsn+ or N:fpg+
 - homonyms
tumor: tissue or disease
sto: table or hundred

Disambiguating corpora

- Manual disambiguation
 - not errorless (75% of agreement!)
 - time and money consuming
 - not widely available
- Automatic disambiguation
 - not errorless
 - more efficient
 - consistent
 - tools not widely available

Automatic disambiguation

- Local grammars
 - model local agreements within a phrase and discard non-applicable interpretations
 - examples in Serbian (Nenadic & Vitas, 1998)
 - ♦ agreement in case, number and gender within NPs
 - ♦ agreement within prepositional phrases (25% text)
- Statistical approaches
 - model stochastic features and correlations
 - example in Serbian
 - ♦ selecting the most frequent interpretation is correct in 95% (Ilic & Kostic, 2002)

Learning from corpora

- Goal: automatically customise existing NLP systems to a new domain or to a new language by learning relevant constraints
- Different targets/linguistic information to learn
 - lexical knowledge/constraints (e.g. entities for IE)
 - syntactical knowledge/constraints (e.g. patterns)
 - semantic knowledge/constraints (e.g. IR designators)

A sample of tagged text

Na <CN type="education">Filozofskom fakultetu u Beogradu</CN> svi profesori koji nisu potpisali ugovore o radu, tridesetak njih dobili su u petak popodne pismenu odluku dekana kojom se svi rasporedujemo na poslove i radne zadatke u <CN type="education">Centru za naučno-istraživački i publicistički rad</CN> u vremenu od 7 i 30 do 15 sati. Ovakav centar je nepostojeci na <CN type="education">Filozofskom fakultetu</CN>.

...

BEOGRAD - Suspendovani profesori <CN type="education">Pravnog fakulteta</CN> podeljeni su u dve grupe, polovinu plate primaju oni koji imaju izdržavane članove porodice, a trecinu ostali. Postojala je inicijativa sa vishe strana da se osnuje fond solidarnosti, ali mi za sada nismo za tu varijantu, iako ne isključujemo mogućnost da se to ucini na nivou <CN type="education">Univerziteta</CN> ili od strane <CN type="education">Udruženja nastavnika Istraživača</CN> - rekao je Juce Jovica Trkulja, suspendovani profesor <CN type="education">Pravnog fakulteta</CN>.

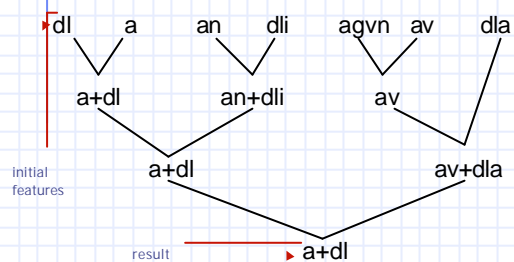
Learning grammatical constraints

- Goal: learn cases "required" by a specific preposition for NPs within PPs (in Serbian)
- Examples:
 - od <NP:genitive>
 - na <NP:accusative or dative or locative>
- NP cases within PPs are already known, but can we learn them from non-disambiguated corpora?

Method

- Determine intersection of features, i.e. minimal set of morpho-syntactic features inherent for every NP within a specific PP type
- Genetic algorithm approach is used to automatically calculate the intersection
- Case tags (obtained from initially tagged corpus) are crossed between pairs of NPs

Learning cases: a sample



An example of crossover between the case tags for PPs containing preposition *na* (Eng. *on*)

Learning cases: results

Preposition	Initial features sample	Learned features	Theoretical features
prema (Eng. <i>towards</i>)	dl dli dlj	dative or locative	dative or locative
od (Eng. <i>from</i>)	ga g gnv ng gnv	genitive	genitive
na (Eng. <i>on</i>)	dl a an dli agvn av dla	accusative or dative or locative	accusative or locative

Terminology acquisition

- Knowledge encoded in textual documents is characterised by sets of terms
- Term = linguistic realisation of a specialised concept in a specialised domain
 - protein kinase C, nuclear hormone receptor
 - real-time computing, DES algorithm
- Applications
 - assigning terms to documents for semantic document categorisation and IR
 - term-based knowledge acquisition/mining, IE
 - constructing terminologies, ontologies, taxonomies

Terminology acquisition

- ✦ Index terms vs. technical terms
 - index terms allow discrimination between documents
 - technical terms indicate semantic concepts in specialised sub-languages
- ✦ Index terms are not necessarily technical terms
 - a set of index terms typically include a mixture of general language words and technical terms
- ✦ Can we extract terms automatically?

Terminology acquisition: method

- ✦ C/NC-value method (Ananiadou & Frantzi, 2000) for automatic extraction of multiword terms
- ✦ Domain-independent, hybrid approach:
 - Linguistic term formation patterns
 - ♦ local grammars
 - Statistical information
 - ♦ frequency of occurrence, "nestedness", length
 - Contextual information
 - ♦ frequent words appearing with term candidates
- ✦ Output:
 - list of terms ranked according to their termhoods

Experiments in English

- ✦ Corpora in biomedicine, crop science, newswire
- ✦ Texts are initially tagged by the EngCG tagger (Voutilainen & Heikkilä, 1993)
- ✦ No further processing
- ✦ Results on a Medline corpus (2082 abstracts)
 - precision: 96% for top-ranked terms

Sample of biomedical terms

<u>retinoic acid receptor</u> retinoic acid receptor retinoic acid receptors RAR, RARs	236.33
<u>nuclear receptor</u> nuclear receptor nuclear receptors NR, NRs	216.00
<u>nuclear receptor co repressor</u> nuclear receptor co-repressor NCoR	214.75
<u>all-trans retinoic acid</u> all trans retinoic acid all-trans-retinoic acids ATRA, at-RA	214.25

Experiments in Serbian

- ✦ Texts are initially tagged by e-dictionaries (Vitas, 1993)
- ✦ No disambiguation techniques are applied
- ✦ Experiments with samples from textbooks in maths and computer science (120k words)
- ✦ Filters based on structure of NPs (Nenadic & Vitas, 1998)
- ✦ Conflating morphological variations only
- ✦ Precision: 94% for top-ranked terms

Sample of mathematical terms

metricxki prostor	633.55
topolosxki prostor	175.13
otvoren skup	93.20
normiran prostor	88.00
Kosxijev niz	68.11
zatvoren skup	59.20
vektorski prostor	53.13
nejednakost trougla	33.98
Hausdorfov topolosxki prostor	19.43
Lagranzova teorema o srednxoj vrednosti	2.32

Sample of CS terms

funkcionalna zavisnost	156.08
visheznacna zavisnost	129.32
skup atributa	81.41
uslov integriteta	40.87
strani klxucx	37.00
primarni klxucx	35.16
sadrzajxaj relacijex	16.35
zavisna relacija	16.00
logixcko projektovanxex baze podataka	10.33
kaskadno brisanxex	8.17
prirodno spajanaxex	8.17

Conclusions

- ❖ Three case studies
- ❖ Lexical acquisition
 - extracts **union** of all word sequences that match a manually defined general LG
- ❖ Learning grammatical constraints/patterns
 - extracts minimal **intersection** of morpho-syntactic features that are inherent to all entities
- ❖ Terminology acquisition
 - combine general LGs for extracting candidate terms, and **statistics** for estimation of their significance

Current and further research

- ❖ Lexical acquisition
 - learning frozen, multiword adverbial expressions
 - designators for another language? (Greek)
- ❖ Constraints learning
 - learn verb selectional preferences in a specific domain through supervised learning (ontology as a seed)
- ❖ Terminology-based translation
 - C/NC-value applied on parallel corpora
 - are term rankings consistent across languages?
 - currently support for English, Greek and Serbian