Comparing Ontology-based and Corpus-based Domain Annotations in WordNet.

A paper by:

Bernardo Magnini

Carlo Strapparava

Giovanni Pezzulo

Alfio Glozzo

Presented by:

rabee ali alshemali

Motive.

Domain information is an emerging topic of interest in relation to WrodNet.

Proposal

An investigation into comparing and integrating ontology-based and corpus-based domain information.

WordNet Domains

- (Magnini and Cavaglia 2000).
- An extension of WordNet 1.6
- Provides a lexical resource, where WordNet synsets have been manually annotated with domain labels, such as: Medicine, Sport, and Architecture.
- The annotation reflects the lexico-semantic criteria adopted by humans involved in the annotation and takes advantage of existing conceptual relations in WordNet.

Question!

- How well this annotation reflects the way synsets occur in a certain text collection ??

Why is this important?

- It is particularly relevant when we want to use manual annotation for text processing tasks (e.g. Word Sense Disambiguation.)

Example to Illustrate:

- Consider the following synset:

{heroin, diacetyl morphine, horse, junk,scag, smack}.

- It is annotated with the Medicine domain because heroin is a drug, and that is maybe best described as medical knowledge.

Example to Illustrate: Cont.

- On the other hand (on the text side), if we consider a news collection – Reuters corpus for example – the word heroin is likely to occur in the context of either:

✓ Crime news.

✓ Administrative news.

And without any strong relation with the medical field.

The moral behind the example:

- We can clearly see the difference:

- Manual annotation considers the technical use of the word.

- Text, on the other hand, records a wider context of use.

How to reconcile?

- Both sources carry relevant information, so supporting ontology-based domain annotations with corpus-based distribution will probably give the best potential for content-based text analysis.

What is needed?

- First Step: a methodology is required to automatically acquire domain information for synsets in WordNet from a categorized corpus.

- Reuters corpus is used because it is free and neatly organized by means of topic codes, which makes comparisons with WorldNet domains easier.

Optimal Goal

- A large-scale automatic acquisition of domain information for WordNet Synsets

However,

- The investigation was limited to a small set of topic codes.

Why is domain information interesting?

- Due to its utility in many scenarios such as:

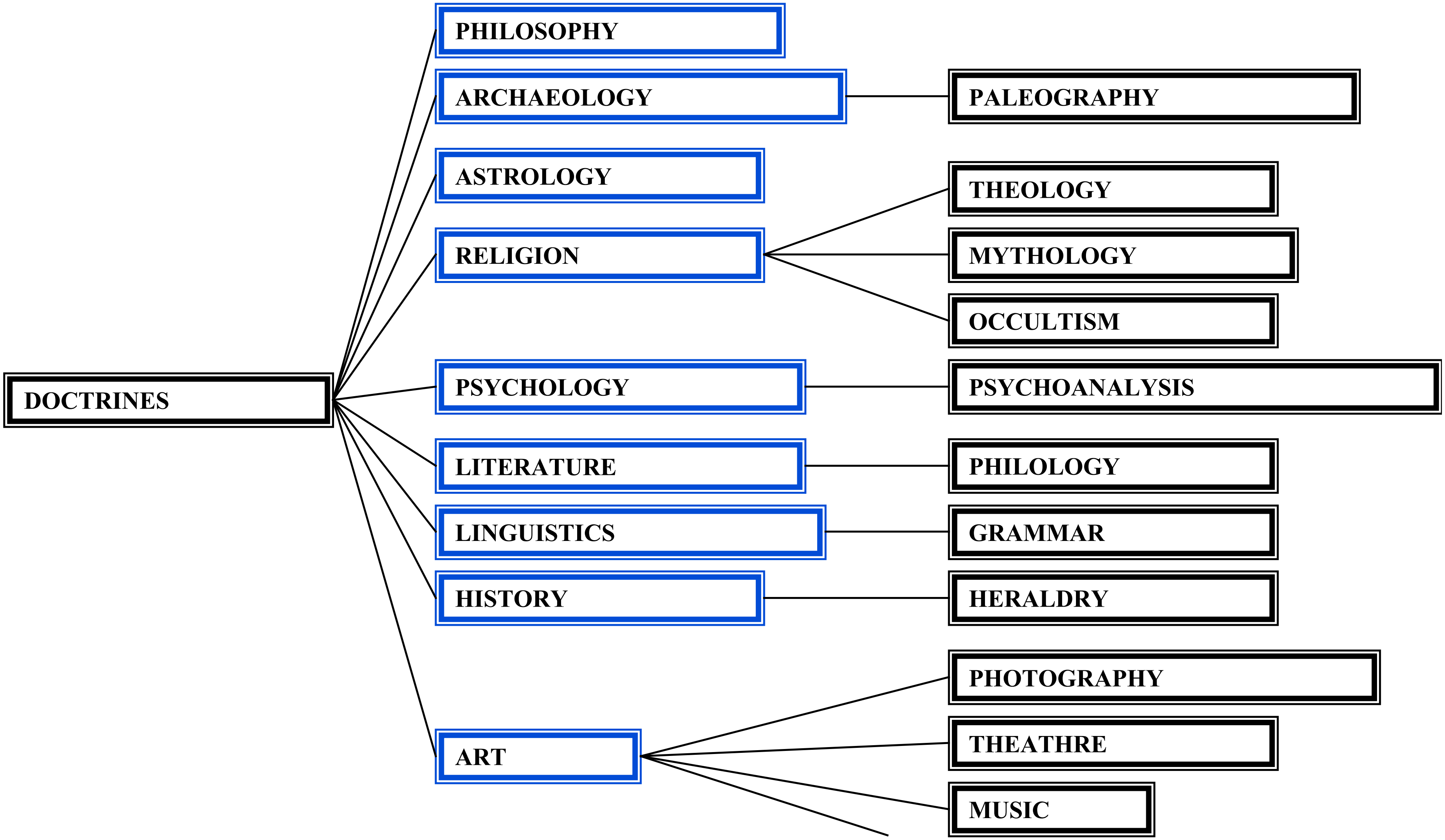
➢ Word Sense Disambiguation (WSD): where information from domain labels are used to establish semantic relations among word senses.

➢ Text Categorization (TC): Where categories are represented as symbolic labels.

WordNet Domains.

- Domains have been used to mark technical usages of words.
- In dictionaries, it is used only for a small portion of the lexicon. Therefore:
- WordNet Domains is an attempt to extend the coverage of domain labels with an already existing lexical database.
- WordNet (version 1.6) Synsets have been annotated with at least one domain label selected from a set of about 200 labels hierarchically organized.

WordNet Domains



- DOCTRINES
  - PHILOSOPHY
  - ARCHAEOLOGY
    - PALEOGRAPHY
  - ASTROLOGY
  - RELIGION
    - THEOLOGY
    - MYTHOLOGY
    - OCCULTISM
  - PSYCHOLOGY
    - PSYCHOANALYSIS
  - LITERATURE
    - PHILOLOGY
  - LINGUISTICS
    - GRAMMAR
  - HISTORY
    - HERALDRY
  - ART
    - PHOTOGRAPHY
    - THEATHRE
    - MUSIC

WordNet Domains.

- Information brought by domains is complementary to what is already in WrodNet.

Three key Observations:

1- A domain my include synsets of different *syntactic categories*, For example:

The medicine domain groups together senses from Nouns such as doctor#1, and hospital#1, and also from Verbs, such as operate#1.

WordNet Domains

2- A domain may include senses from different WordNet sub-hierarchies, for example:

The sport domain contains senses such as:

-- Athlete#1, from life_form#1

-- game_equipment#1, from physical_object#1

-- sport#1, from act#2

-- playing_field#1, from location#1

WordNet Domains.

3- domains may group senses of the same word into homogenous clusters, but:

side effect → Reduction in word polysemy.

WordNet Domains.

- The word “bank” has 10 different senses.
- Three of them (#1, #3, and #6) can be grouped under the Economy domain.
- While #2 and #7 both belong to the Geography and Geology domain.
- → Reduction of the polysemy from 10 to 7 senses.

Sense	Synset and Gloss	Domains
#1	Depository financial institution, bank, banking, banking company.	Economy
#2	bank (sloping land …)	Geography, Geology
#3	bank (a supply or stock held in a reserve)	Economy
#4	bank, bank building (a building …)	Architecture, Economy
#5	bank, (an arrangement of similar objects.	Factotum
#6	savings bank, coin bank, money box.	Economy
#7	bank, (a long ridge or pile…)	Geography, Geology
#8	Bank (the funds held by a gambling house …)	Economy, Play
#9	bank, cant camber (a slope in the the turn of a road …)	Architecture
#10	bank (a flight maneuver…)	Transport

Procedure for synset annotation.

- It is an inheritance-based procedure to automatically mark synsets
- A small number of high level synsets are manually annotated with their pertinent domains
- An automatic procedure exploits WrodNet relations (i.e. hyponymy, antonymy, meronymey…) to extend the manual assignments to all reachable synsets.

Example.

- Consider the following synset:

  {beak, bill, neb, nib}

- It will be automatically marked with the code Zoology, starting from the synset {bird} and following "part_of" relation.

Issues!

Oh man!, why there always have to be issues !? :o)

➢ Wrong propagation. Consider:

barber_chair#1 is "part_of" barber_shop#1

barber_shop#1 is annotated with Commerce

→ barber_chair#1 would wrongly inherit the same domain.

✓ Therefore, in such cases, the inheritance procedure has to be blocked to prevent wrong propagation.

How to fix …

- The inheritance procedure allows the declarations of "*exceptions*"
- Example:

  Assign shop#1 to Commerce

  With exception[part, isa, shop#1]

which assigns the synset shop#1 to Commerce, but excludes the parts of the children of shop#1 such as barbershop#1.

Issues. Cont.

- FACTOTUM: a number of WordNet synsets do not belong to a specific domain, but can appear in many of them; Therefore, a *Factotum label* is created for this purpose.
- It includes two types of synsets:

1- Generic synset.

2- Stop sense synsets.

Generic Synsets.

- They are hard to classify in a particular domain.
- Examples:

Man#1 : an adult male person (vs. woman)

Man#3 : any human being (generic)

Date#1 : day of the month.

Date#3 : appointment, engagement.

- They are placed high in the hierarchy – many verb synsets belong to this category –

Stop Sense Synsets.

- Include non polysemous words.
- Behave as stop words since they don’t contribute to overall sense of text.
- Examples:

  Numbers, Weekdays, colors …

Specialistic vs. Generic Usages.

- About 250 domain labels in WordNet Domains.
- Some synsets occur in well-defined context in the WordNet hierarchy, but have a wider (generic) *textual* usage.
- Example:

The synset {feeling} -- the psychological feature of experiencing affective and emotional states.

✓ It could be annotated under Psychology domain.

✓ the use of it in documents is broader than the psychological discipline.

→ a Factotum annotation is more coherent.

Corpus-Based Acquisition procedure

- Automatically acquire domain information from the Reuters corpus and compare it with domain annotations already present in WrodNet domains.
- Steps:

1- Linguistic Processing of the corpus.

2- acquisition of domain information for WordNet synsets based on probability distribution in the corpus.

3- Matching of required information with domain manual annotations.

Experimental Setting.

- Reuters corpus has about 390,000 English news.
- Each one is annotated with at least one topic code.
- Only limited subset of the codes were considered.

Domain	Topic codes	# Reuters tokens
Religion	GREL	307219
Art	GENT	400637
Military	GVIO	3798848
Law	GCRIM	2864378
Sport	GSPO	2230613

Linguistic Processing.

- The subset of Reuters corpus was first *lemmatized* and annotated with part of speech tags.
- WordNet morphological analyzer was used to resolve ambiguities and lemmatization mistakes
- A filter was applied to identify the words actually contained in WordNet 1.6
- The result is 36,503 lemmas including 6,137 multiwords.

Acquisition Procedure.

- Given a synset in WordNet Domains.
- Need to identify which domain, among the ones selected for the experiment, is relevant in the Reuters corpus.
- *A relevant Lemma list* for a synset is built as the *union* of the synonyms and of the content words of the gloss for that synset.
- The list represents the context of the synset in WordNet, and is used to estimate the probability of a domain in the corpus.
- The probability is collected in a Reuter Vector, with one dimension for each domain.
- The value of each dimension is the probability of that domain.
- The probability of the synset for a domain is conditioned by the probability of its most related lemmas.
- I am not gonna include the equations here … :o)

Matching with Manual Annotation.

- In addition to the Reuters vector, a WordNet Vector is built for each synset with a dimension for each selected domain.
- The selected domains gets a score of 1; others gets a score of 0.
- The two vectors are normalized
- The scalar product is computed for the two vectors.
- What we get is a *proximity score* between the two sources of domain information.
- The score ranges from 0 → 1 and indicates similarity between the two annotations.

Experiment 1: Synsets with *unique* manual annotations.

- Two restrictions applied:

✓ a synset must have at least one word among its synonyms occurring at least once in the Reuter corpus.

✓ It must have just one domain annotation in WordNet domains.

- This selection produced 867 experimental synsets.
- Average proximity score was very high (0.96) indicating a very relevant subset of synsets.

Example.

- The synset: {baseball, baseball game, ball game – (a game played with a bat and ball between two teams of 9 players; teams take turns at bat trying to score run)}
- It was manually annotated with the Sport domain.
- WordNet vector shows 1 for Sport, 0 elsewhere.
- The procedure produced the following vector:

Law	Art	Religion	Sport	Military
1.82^{e-60}	2.44^{e-55}	1.71^{e-152}	1	2.45^{e-63}

Experiment 2: Synsets with multiple manual annotations.

- A number of synsets where annotated with multiple domain labels in WordNest domains.
- Example: consider the synset of the adjective canonic#2 :{canonic, canonical – (of or relating to or required by cannon law)}
- It’s annotated with two labels: *Religion*, and *Law*.
- Corresponding Reuter’s vector:

Law	Art	Religion	Sport	Military
0.41	**9.48^{e-47}**	**0.56**	0.004	**0.02**

Experiment 3: Factotum Annotations.

- Factotum synsets don’t belong to any specific domain.
- Should have high frequency in all the Reuters texts.
- Example:

The synset containing the verb “to be” {be – (have the quality of being)}, corresponds to the following Reuter vector.

Law	Art	Religion	Sport	Military
0.21	**0.29**	**0.20**	0.16	**0.20**

Experiment 4: Mismatching Annotations.

- For some synsets, the WrodNet vector and Corpus vector produced contradictory results.
- Exmaple: consider the synset {wrath, anger, ire, ira – (belligerence aroused by a real or supposed wrong (personified as one of the deadly sins))}
- It is annotated with Religion, inherited from its *hypernym* {moral sin, deadly sin}.
- Its Corpus vector is:

Law	Art	Religion	Sport	Military
1.4^{e-45}	3.5^{-44}	5.2^{-13}	9.48^{-48}	1

- Reason: Military nature of most of the lemmas, and the fact that the only Religious lemma {deadly sin} is rare in Reuters corpus.

Experiment 5: Covering problems.

- The relevant lemma list for some synsets are not well covered in the Reuters corpus
- Example: the synset {Loki – (trickster; god of discord and mischief; contrived death of Balder and was overcome by Thor)}. Which is manually annotated with *Religion*, due to its *hypernym* {deity,divinity, god, immortal}.
- Its Reuters vector is:

Law	Art	Religion	Sport	Military
2.10^{e-44}	1.45^{-131}	2.63^{-13}	6.78^{-68}	**1**

- The preferred domain Military depends on the absence, in the corpus of lemmas such as (Loki, Balder, Thor) and the presence of military lemmas such as (discord, death, overcome).

Summary and Conclusions.

- We have looked at:

o WordNet Domains as a lexical resource.

o Procedure for automatic acquisitions of domain information.

➢ Ontology-based and corpus based annotations play complementary roles and its difficult to find a mapping between them.

Future work.

- A full automatic procedure for the acquisitions of domain information from corpora.
- Collect and use large and diverse domain annotated corpora.
- The integration of corpus-based domain information with WordNet taxonomy.

Questions?