Automatically Constructing a Wordnet for Dutch

Tim Van de Cruys

RCEAL, University of Cambridge

CLIN 21
11 February 2011
Ghent

Introduction

- Manual construction of a semantic hierarchy is a tedious and time-consuming job
- Automatic methods for lexico-semantic information extraction provide accurate results
  - Extraction from semi-structured resources (Wikipedia, Wiktionary, . . .)
  - Distributional similarity
  - Translation dictionaries from automatically aligned parallel corpora
- Goal: combine various methods in order to automatically construct a wordnet for Dutch

Introduction

- Synset structure of original Princeton Wordnet 3.0 for English is used as base structure
- *extend approach*: goal is to augment Princeton Wordnet synsets with correct Dutch translations

Mining Wiktionary

- English Wiktionary contains translations into Dutch, divided by sense
- Wiktionary senses are automatically mapped to Princeton Wordnet senses using the **definition glosses** of both resources
- Dutch translations are assigned to resulting synsets

Sense mapping

- COLD (wiktionary senses)
  - A condition of low temperature
  - (medicine) A common, usually harmless, viral illness, usually with congestion of the nasal passages and sometimes fever
- COLD (wordnet synsets)
  - a mild viral infection involving the nose and respiratory passages (but not the lungs)
  - the sensation produced by low temperatures

Sense mapping

- COLD (wiktionary senses)
  - A condition of low temperature
  - (medicine) A common, usually harmless, viral illness, usually with congestion of the nasal passages and sometimes fever
- COLD (wordnet synsets)
  - a mild viral infection involving the nose and respiratory passages (but not the lungs)
  - the sensation produced by low temperatures

Sense mapping

- COLD (wiktionary senses)
  - A condition of low temperature
  - (medicine) A common, usually harmless, viral illness, usually with congestion of the nasal passages and sometimes fever
- COLD (wordnet synsets)
  - a mild viral infection involving the nose and respiratory passages (but not the lungs)
  - the sensation produced by low temperatures

Sense mapping

- COLD (wiktionary senses)
  - A condition of low temperature → *kou, koude*
  - (medicine) A common, usually harmless, viral illness, usually with congestion of the nasal passages and sometimes fever → *verkoudheid*
- COLD (wordnet synsets)
  - a mild viral infection involving the nose and respiratory passages (but not the lungs)
  - the sensation produced by low temperatures

Sense mapping

- COLD (wiktionary senses)
  - A condition of low temperature → *kou, koude*
  - (medicine) A common, usually harmless, viral illness, usually with congestion of the nasal passages and sometimes fever → *verkoudheid*
- COLD (wordnet synsets)
  - a mild viral infection involving the nose and respiratory passages (but not the lungs) → *verkoudheid*
  - the sensation produced by low temperatures → *kou, koude*

Alignment-based extraction

- Wordnet extraction based on parallel corpora (Sagot and Fišer 2008)
- Translation dictionaries can straightforwardly be extracted from automatically aligned parallel corpora
- Problem: no discrimination among different senses
- Combine with distributional and wordnet-based similarity measures in order to determine the correct sense

example: EN *cold*

- Look up translations for EN *cold* → NL *kou, verkoudheid*
- Find NL similar words (syntax-based distributional similarity model)
  - *kou*: 'koude', 'vrieskou', 'hitte', 'regen', 'winterkou'
  - *verkoudheid*: 'bronchitis', 'griep', 'keelontsteking', 'hooikoorts', 'griepje'
- Translate NL similar words back into EN
  - *kou*: 'cold', 'heat', 'heatwave', ...
  - *verkoudheid*: 'bronchitis', 'flu', 'influenza', ...
- Compute wordnet-based distance between synsets of EN *cold* and EN similar word translations

example: EN *cold*

- average pairwise wu & palmer similarity between 'low temperature' synset i_1 and EN similar words to *kou*

$$\frac{\sum_{j \in c_{kou}}^{n} sim_{wup}(i_1, j)}{n} = 0.97 \quad (1)$$

- average pairwise wu & palmer similarity between 'low temperature' synset i_1 and EN similar words to *verkoudheid*

$$\frac{\sum_{j \in c_{verkoudheid}}^{n} sim_{wup}(i_1, j)}{n} = 0.18 \quad (2)$$

example: EN *cold*

- Likewise, average pairwise wu & palmer similarity between 'infection' synset i_2 and EN similar words to *kou*

$$\frac{\sum_{j \in c_{kou}}^{n} sim_{wup}(i_2, j)}{n} = 0.50 \quad (3)$$

- average pairwise wu & palmer similarity between 'infection' synset i_2 and EN similar words to *verkoudheid*

$$\frac{\sum_{j \in c_{verkoudheid}}^{n} sim_{wup}(i_2, j)}{n} = 0.92 \quad (4)$$

Alignment-based extraction

- Distributionally similar words (translated back into English) combined with wordnet-based similarity measure allows for mapping to correct synset
- Wordnet-based similarity threshold is used to select correct sense assignments

Introduction

- Hypernym/hyponym extraction usually based on some form of 'Hearst patterns'
  - **fruits** such as *apples* and *bananas*
  - *kumquats*, *pomegranates* and other exotic **fruits**
- Effective but noisy
- Combination with large word cluster database allows for filtering noisy hypernym/hyponym combinations (Van Durme and Pasca 2008)

Method

- Patterns used:
  - subject – nominal predicative complement dependencies from first sentences of Wikipedia
    - *xylofoon*#PREDC_N#**muziekinstrument**
  - appositions
    - De Franse **president** *Nicolas Sarkozy*
- Word Clustering:
  - Large 200K word clustering based on $\pm$ 1M syntactic features, clustered into 2K clusters
  - Constraints with regard to clustering
    - a particular hypernym needs to cover at least a ratio σ of the total number of instances of a cluster (i.e. a hypernym should be general enough)
    - a particular hypernym must not cover more than a ratio τ of all clusters (but not too general)

example: *muziekinstrument*

- **muziekinstrument**: *trekzak, klokkenspel, vibrafoon, hoorn, synthesizer, snaarinstrument, berimbau, contrabas, xylofoon, keyboard, accordeon, sampler, marimba, bas, piano*
- **vis**: *zeelt, Grondel, roodbaars, marlijn, griet, snoekbaars, schol, vis, baars, kopvoorn, steur, zalm, zeebaars, stekelbaars, beekforel, spiering, heilbot, zeewolf, zeeduivel, ansjovis, zeebrasem, zwaardvis, brasem, tong, poon, koolvis, haring, wijting, riviergrondel, blankvoorn, kabeljauw, paling, Heek, makreel, tonijn, alver*
- **munteenheid**: *roebel, won, peso, lira, yen, zloty, roepie, baht, peseta, escudo, pond, lire, gulden, euro, shilling, frank, dollar, som, mark*

Implementational details

- Extraction from semi-structured resources:
  - Wiktionary parser implemented in Python
- Extraction from translation dictionaries
  - Word alignment database from OPUS corpus (Tiedemann 2009)
  - Word similarity database based on TWNC/MEDIARGUS (2 billion words) – parsed with ALPINO for dependency triples
- Hyponym/hypernym extraction
  - First sentences from Wikipedia + appositions from TWNC/MEDIARGUS
  - *k*-means clustering (with CLUTO) using TWNC/MEDIARGUS
- Wordnet format:
  - Conversion of Wordnet 3.0 to RDF (VU)
  - Output as RDF triples – extension to Wordnet RDF

Results

- Focus on nouns
- 11008 synsets are filled with at least one Dutch translation
- 11783 nouns added in total, accounting for 17182 senses
- Original wordnet: 82115 nouns, 146347 senses

Evaluation

- Manual evaluation of 200 randomly selected synsets
- For each synset:

$$Precision = \frac{\#correct\ \ noun\ \ senses}{all\ \ noun\ \ senses} \tag{5}$$

- Average precision of added noun senses: **0.82**
- 2935 new nouns added that are not in CORNETTO

New additions

boerenzwaluw, hardheid, euro, vergevingsgezindheid, wasdraad, epoch, onbepaalde wijs, caracal, duivenkot, allusie, stormvogel, stuurprogramma, apennoot, arachidenoot, burgerlijke bouwkunde, ransuil, kamper, selenologie, paleolithicum, jaarlijkse uitgave, jaaruitgave, haarkapper, koalabeertje, opportuniste, continentaal plat, waterlelie, bosuil, plaatsigheid, ariteit, xyleem, tandenfee, nirwana, thijm, bloes, meidoren, Maleier, koalabeer, ongewenstheid, coördinatenstelsel, open veld, stamreeks, kwartierstaat, stamboomonderzoek

Conclusion

- Using a number of simple extraction algorithms, a Dutch Wordnet can be constructed automatically with great precision
- Combination of different techniques is able to overcome the noisiness inherent to the individual algorithms
- Still work to do with regard to recall

Future work

- Combine techniques described here with other known methods for wordnet extraction
  - Use of Wikipedia's inter-language links (Ponzetto & Navigli 2009)
- Combine algorithm for hypernym extraction with hierarchical clustering techniques
- Use of distributional similarity to improve on sense mapping for Wiktionary translations
- Automatic evaluation using CORNETTO