Attribute Value Acquisition Through Clustering of Adjectives

Agnieszka Mykowiecka Małgorzata Marciniak

Institute of Computer Science, Polish Academy of Sciences
{agn,mm}@ipipan.waw.pl

PolTAL, September 2014

Domain Model Creation

- concept recognition
  - terminology extraction methods

    left kidney *acute appendicitis* *capital gains tax*
- relation between concepts
  - hypernymy/hyponymy

    capital gains tax is a tax ...
  - meronymy, relative locations, time precedence, reason ...

    The superior border of the right kidney is adjacent to the liver
- attribute recognition

  location type size color
- attribute value recognition

  left acute

Experiment

Goal

Automatically identify and group attribute values by an analysis of Polish descriptive adjectives which occur in domain related texts.

Method

- selection of descriptive adjectives from text,
- clustering adjectives on the basis of a set of context related features and interword relations from plWordnet.

Previous works

- using relational adjectives for extraction of hyponyms from medical texts (Acosta 2013);
- using adjectives for attribute learning (Almuhareb and Poesio 2004) and (Cimiano 2006);
- a semi-supervised machine-learning approach for the classification into property denoting vs. relation denoting adjectives (Hartung and Frank 2010);
- corpus-based methods used to group Polish adjectives into semantic clusters (Piasecki et al. 2009);
- clustering adjectives as the first stage of automatic identification of adjectival scales (Hatzivassiloglou and McKeown 1993).

Classification of Adjectives

- traditionally, adjectives are divided according to a difference in the type of denoted properties:
  - qualitative (i.e. absolute, descriptive)
    niski poziom 'low level'
  - relative (i.e. classifying, distinguishing)
    poradnia alergologiczna 'allergology clinic';
- BEO classification (Boleda 2006):
  - **b**asic (i.e. qualitative, property-denoting) adjectives (*old box*),
  - **e**vent-related adjectives (*eloquent person*),
  - **o**bject-oriented (relational) adjectives (*environmental science*);
- a very detailed classification of Polish adjectives into 56 classes made from the MT point of view (Jassem 2002).

Descriptive and Classifying Adjectives in Polish

- descriptive adjectives in Polish occur typically before nouns
  niski poziom 'low level',
- classifying adjectives are located after a noun
  $jama_{subst}$ $brzuszna_{adj}$ 'abdominal cavity'
  but:
  - lexicalized connections
    biały szum 'white noise' or *ostry dyżur* 'ER',
  - two classifying adjectives
    wojewódzka poradnia alergologiczna
    'provincial allergology clinic'
    poradnia wojewódzka 'provincial clinic'
    poradnia alergologiczna 'allergology clinic'

Data Characteristics

- hospital discharge documents gathered at one of the Polish children's hospitals;
- six departments (general, surgery, neurology, neonatology, infectious diseases and rehabilitation); several physicians;
- anonymised original MS Word files converted into plain text;
- morphologically tagged using Pantera (Acedański 2011) based on the Morfeusz analyser (Woliński 2006);
- the data set consists of 3,116 documents comprising 1,940,000 tokens;
- out of dictionary forms (11,777 in total) tagged as *ign*.

POS Distribution

	POS	types	occurrences
acronym	acron	1,476	25,814
abbreviation	brev	168	76,058
gerund	ger	444	13,405
substantive	subst	3,642	362,352
adjective	adj	1,743	130,611
past participle	ppas	291	23,278

Nominal Phrases with Adjectival Modifiers

Shallow grammars recognized nominal phrases with adjectival modifiers which obeyed standard Polish gender, case and number agreement constraints.

	nouns		length=2		length=3		length>3	
	types	occ.	types	occ.	types	occ.	types	occ.
AN	1,209	22,425	3,323	21,037	654	1,212	94	176
NA	1,070	63,857	3,067	54,510	754	6,349	375	2,998
ANA	310	4,355	—	—	906	3,426	408	929

- Selected adjectives occurred at least 10 times as left modifiers and 2 times more frequently as left than as right modifiers.

Similarity Features

- Lexical Context
- Common Modified Nouns
- Wordnet Similarity

Similarity Features, Lexical Context

- Sequences of base forms of 1, 2 and 3 tokens
- Sequences of POS tags of 2, 3, 4 tokens
- The base form of the nearest:
  - verb
  - noun type token (nouns, gerunds, acronyms, abreviations)
  - adjective type token (adjectives, participles)
  - preposition.
- smilarity defined as Jaccard coefficient on form frequencies

$$sim_t(a_i, a_j) = \frac{number\ of\ all\ occurrences\ of\ common\ C_t\ type\ contexts}{number\ of\ occurrences\ of\ all\ C_t\ contexts\ of\ both\ adjectives} \quad (1)$$

Common Modified Nouns

Nouns representing concepts with the same attribute occur with the same adjectives expressing it:

mały, powiększony, niewielki 'small, enlarged, slight'
mała zmiana, mały węzeł, niewielki węzeł, niewielki ból

$$sim_{com-nouns}(a_i, a_j) = \frac{number\ of\ commonly\ modified\ nouns}{max\ number\ of\ nouns\ modified\ by\ one\ adjective} \quad (2)$$

Wordnet Similarity

- antonymy – (the most numerous relation) may link adjectives expressing different values of one attribute,
  duży, mały 'big, small'
- hypernymy and relatedness – directly connected with meaning similarity,
  dobry, doskonały 'good, perfect'
- antonymy – words which share the same antonyms are similar
  dobry, niezły 'good, quite nice'

$$sim_{w_{a-X}}(a_i, a_j) = \frac{number\ of\ senses\ in\ relation\ X}{smaller\ number\ of\ senses} \qquad (3)$$

Wordnet Similarity, 2

- additional information from the noun hyperonymy hierarchy for nouns automatically obtained via transformation of adjectives into possible noun forms

intensywny, nasilony 'intense' 'severe'
intensywność, nasilenie 'intensity' 'severity'

$$sim_{Wn-hyp}(a_i, a_j) = \frac{\sum_{si \in S(n(a_i)), sj \in S(n(a_j)):\, ((si \ \ hypernim \ \ sj) \, or \, (sj \ \ hypernim \ \ si))} 1}{min\,(S(n(a_i)), S(n(a_j)))} \quad (4)$$

where $n(a_i)$ is a noun derived from a_i.

Data from the Dictionary of Polish Synonyms

- open source dictionary of synonyms (`http://synonimy.ux.pl`), 13,180 groups with 44,550 words or phrases.
- similarity groups for the considered list of adjectives.

$$sim_{dict}(a_i, a_j) = \frac{number\ of\ groups\ with\ both\ elements}{bigger\ number\ of\ groups\ for\ both\ elements} \quad (5)$$

Development Set

- 28 from the most popular adjectives that appeared in AN phrases;
- selection procedure:
  - 105 adjectives that appeared at least 50 times in adjective-noun phrases;
  - 65 those that were used three times more frequently to the left of an adjective than to the right;
  - 28 that create at least two-element groups when clustered according to the information obtained from the plWordnet and the thesaurus of synonyms.

Manual Grouping of 28 Adjectives

1: *drobny* 'small', *niewielki* 'small', *duży* 'large', *powiększony* 'enlarged';
2: *pozostały* 'remaining';
3: *stały* 'stable, constant';
4: *kolejny* 'next', *ponowny* 'repeated', *początkowy* 'initial', *ostatni* 'last', *pierwszy* 'first' *drugi* 'second';
5: *obfity* 'abundant', *liczny* 'numerous', *nieliczny* 'not numerous';
6: *nieznaczny* 'minor, insignificant', *znaczny* 'considerable, substantial', *istotny* 'important';
7: *obustronny* 'two-sided'/'mutual;
8: *rozluźniony* 'loose' 'relaxed';
9: *podwyższony* 'increased', *wzmożony* 'enhanced', *niski* 'low', *wysoki* 'high',
10: *różny* 'different';
11: *płynny* 'liquid'/'floating',
12: *silny* 'strong', *słaby* 'weak'

Clustering

- modified hierarchical clustering; Multidendrograms, complete linkage strategy (Fernandez and Gómez 2008)
- input: singular similarity values for all pairs of adjectives which were counted as a weighted sum of 26 features

$$sim(a_i, a_j) = \sum_{sim_t \in S_{sim}} weight(sim_t) \times sim_t(a_i, a_j) \quad (6)$$

Tunning of Weights – Models Tested

- models with the only one non-zero value
- models with only left or right context
  - F-measure of 0.601 for left and 0.556 for right (all) contexts
  - left and right sets of coefficients based on nearest word:

    combining all nearest left word 0.622 right 0.577

    only left noun 0.609 verb 0.569 adjective 0.565 preposition 0.541

Results for Groups of Coefficient Types

type of similarity	F-measure
left noun/adjective/verb/preposition	0.622
right noun/adjective/verb/preposition	0.577
both nouns/adjectives/verbs/prepositions	0.699
left strings of base forms and pos	0.601
right strings of base forms and pos	0.556
both strings of base forms and pos	0.636
wordnet relations	0.655
common nouns	0.698–0.75

Results for Development Set

- B-cubed measure (Bagga and Baldwin 1998)
- the **best manually tuned** model:
  F-measure of **0.799** for 12 groups, 0.813 for 13 groups.
- If the information from the **thesauri** was **neglected** the F-measure dropped only to **0.758**.

Manually Tuned Model

	coeff.	val.	coeff.	val.	coeff.	val.	coeff.	val.	coeff.	val.
POS										
	lpos2	0.12	lpos3	0.12	lpos4	0.12				
	rpos2	0.09	rpos3	0.09	rpos4	0.09				
base form										
	lb1	0.12	lb2	0.12	lb3	0.12				
	rb1	0.09	rb2	0.09	br3	0.09				
nearest verb, noun, adjective, preposition										
	lvp	0.12	lnp	0.30	ladj	0.25	lpp	0.06		
	rvp	0.09	rnp	0.25	radj	0.20	rpp	0.04		
thesauri										
	a-sim	0.20	a-ant	0.40	a-hyp	0.10	n-hyp	0.15	dict	0.2
common nouns	**0.5**									

Evaluation

- 101 adjectives of frequency >30, 2 times more often in AN than NA position
- 52 (manual) groups, 28 of them containing only one element
- results for the model **manually tuned** on 28 adjectives:
  precision 0.676, recall 0.653 and F-measure: **0.664**
- the model consisting of **equal coefficients**
  precision 0.626, recall 0.620 and F-measure **0.623**

An Exemplary Result

method	set	group				
automatic	101	*dodatni*, positive,	*niski*, low,	*wysoki*, high,	*obniżony*, reduced,	*podwyższony* increased
manual	101		*niski*, low,	*wysoki*, high,	*obniżony*, reduced,	*podwyższony* increased
manual & automatic	28		*niski*, low,	*wysoki*, high,		*podwyższony* increased

Conclusions

- Automatic clustering can be used to preliminary identify adjectives describing one property in big enough and coherent data.
- Adjectives that are not related with other adjectives in plWordnet or other thesauri can be clustered only on the basis of the syntactic information available in the corpus .
- The most informative feature in our model was the coefficient based on common nouns modified by both compared adjectives.
- The left contexts turned out to be only slightly more important than the right ones, and contexts based on the nearest nouns/adjectives/verbs/prepositions are only slightly better than contexts based on POS and base forms.

Further plans

- extending the approach with adjectives' sense disambiguation task using clustering methods allowing for the placement of one word in more than one cluster like in (Tomuro et al. 2007)

Thank you for your attention.