Extraction of Polish multiword expressions

Paweł Chrząszcz

Computational Linguistics Department, Jagiellonian University
ul. Gołębia 24, 31-007 Kraków
Computer Science Department, AGH University of Science and Technology
Aleja Adama Mickiewicza 30, 30-059 Kraków

11th International Workshop on Natural Language Processing and Cognitive Science
Venice, Italy

October 29, 2014

Outline

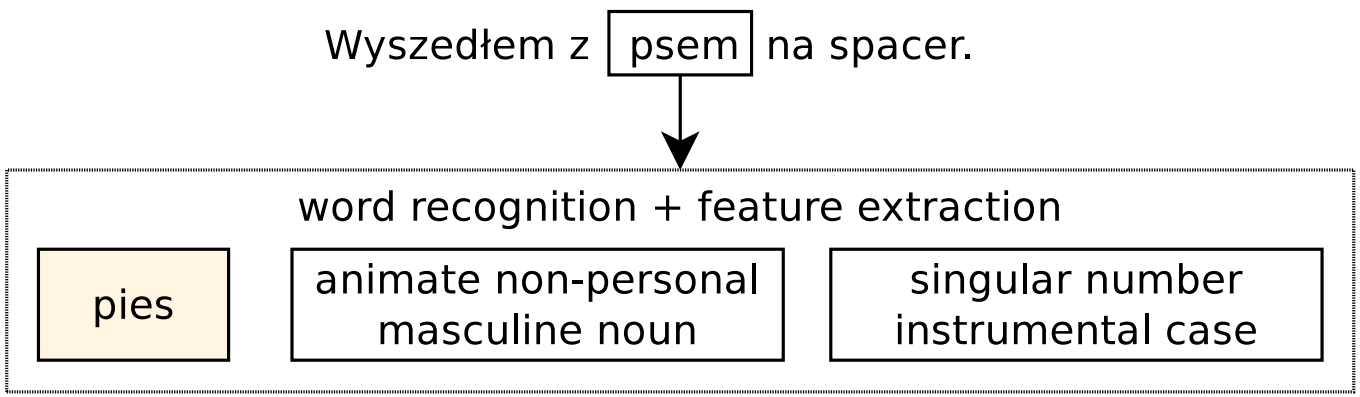
Introduction

Natural language processing of fusional languages usually requires extraction of language-dependent features.

Example

*Wyszedłem z **psem** na spacer.* (*I went for a walk with my dog.*)

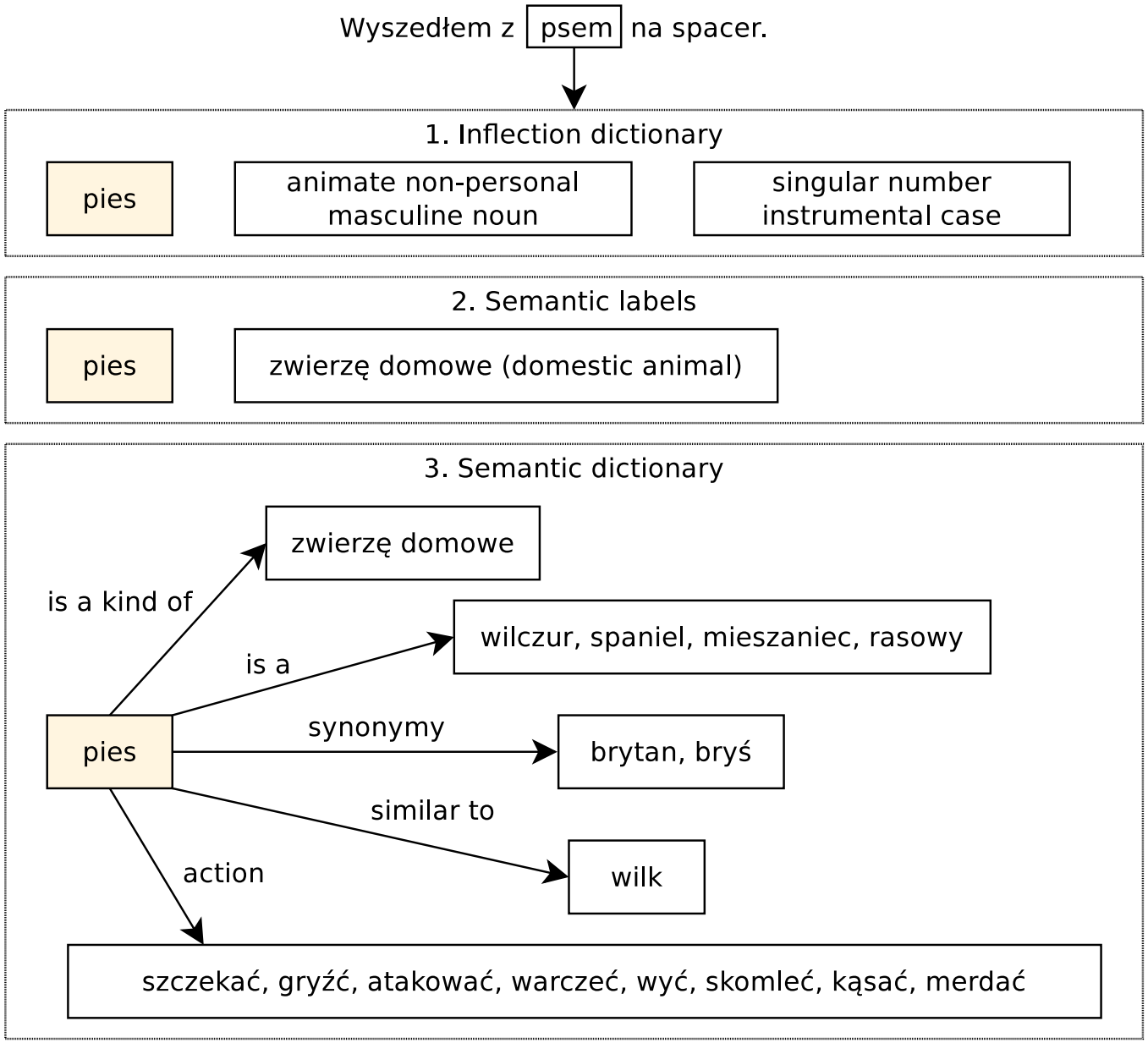
- The token "psem" is recognized as a form of the noun "pies".
- Extracted features are: lemma, gender, number and case.



Inflecting language processing methods

- Features may be extracted using a statistical algorithm, trained on a tagged corpus.
- Morphological analyzers (Morfeusz, Morfologik) provide word recognition function and allow high accuracy tagging of raw text.
- Polish Inflection Dictionary (SFJP) is an alternative to morphological analyzers.
  - For each lexeme (e.g. "zamek" – *castle* and "zamek" – *lock* are two lexemes) it contains a list of inflection forms.
  - It contains morphological relations connecting e.g. verb and its participles (participles are separate lexemes inflected like an adjective), imperfective ("pisać") with perfective ("napisać"), etc.
- The dictionary is still being improved
  - Polish Semantic Dictionary (SSJP) is an extension providing semantic relations for most common words
  - Other words should have short semantic labels

SFJP, semantic labels and SSJP



Need for Multiword Expressions

- SFJP and other linguistic resources for Polish are missing **multiword expressions** (MWEs)
- "idiosyncratic interpretations that cross word boundaries" (Sag et al. [2002])
- They are lexemes that consist of multiple tokens and have properties that differ from what can be inferred from individual tokens
- They have well defined, fixed meaning
- MWEs might have completely new meaning: "panna młoda" means *bride*, while literal meaning would be *young maid*
- Their meaning and inflection patterns should definitely be present in inflection dictionaries to allow their recognition and processing

Our goal

We focus on the extraction of **multiword expressions** (MWEs) from Polish text.

	Words		**MWEs**	
	common	rare	common	rare
Syntax	Inflection dictionaries, morphological analysers, eg. SFJP, Morfologik, Morfeusz	Stemmers, taggers	*MWE dictionary*	*MWE extraction*
Semantics	SSJP, ontologies, WordNet	*Semantic labels*	*Semantic labels*	*Semantic label prediction*

Related work

- **Named Entity** (NE) **Recognition** (NER)
  - **NEs** – phrases belonging to predefined categories: people, geographical objects, dates etc.
  - Extracted using statistical methods (MaxEnt, HMM), rules, gazetteers.
  - They reach F1 of 70-90%, but are limited to selected categories.
- General MWE extraction
  - For some languages, e.g. French, MWEs are included in dictionaries and corpora
  - For other languages, Mutual Information (MI), χ^2, Permutation Entropy – low precision
  - Supervised methods (eg. SVM) with manually tagged samples – F1 up to 90%, but only for binary classification
- No known method of unsupervised MWE extraction for Polish.

Examples of nominal MWEs that we focus on

Word type	**Examples**
Person names	Maciej Przypkowski, Allen Vigneron, Szymon z Wilkowa (*Szymon from Wilków*)
Other proper names	Lazurowa Grota (*Azure Cave*), Polski Związek Wędkarski (*Polish Fishing Association*)
Other named entities	rzeka Carron (*River Carron*), jezioro Michigan (*Lake Michigan*), premier Polski (*Prime Minister of Poland*)
Terms of art	martwica kości (*bone necrosis*), dioda termiczna (*thermal diode*), zaimek względny (*relative pronoun*)
Idioms and other common words	panna młoda (*bride*), piłkarz ręczny (*handball player*), baza wojskowa (*military base*)

Anatomy of Polish nominal MWEs

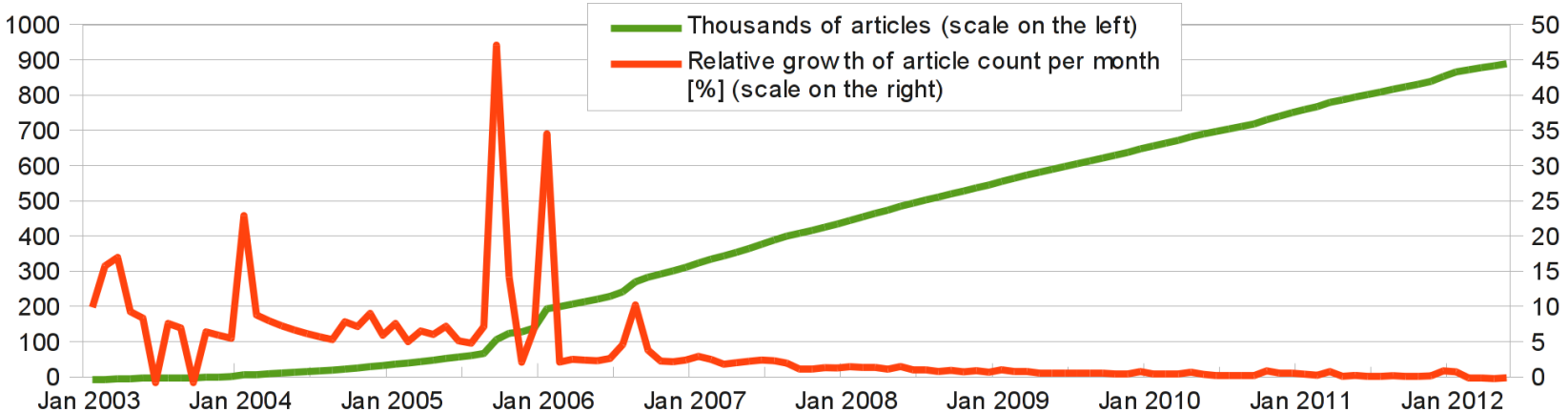
- Nominal MWE: two or more *tokens*: words, numbers or punctuation marks
- A token that is a Polish word can be either fixed or inflectable
- Inflectable words (mostly nouns, adjectives) form the core of the MWE
- Base form: nominative, usually singular

Example: "Związek Lekkoatletyczny Krajów Bałkańskich"

- Meaning: *Athletic Association of Balkan Countries*
- Instrumental case: "Związkiem Lekkoatletycznym Krajów Bałkańskich"

Wikipedia as a source of Polish MWEs

- It is difficult to extract MWEs from a plain text corpus.
- Wikipedia is a much better choice (open, well-stuctured, large)
- Wikipedia is widely used for natural language processing.
  - It may be better than WordNet for word sense disambiguation.
  - Ontologies are created from Wikipedia: YAGO, DBPedia.
  - Article content, infoboxes, page categories and links between articles are used as source data.
  - Crossligual Correspondence Asymmetries might be used to extract nominal semantic non-decomposable MWEs (F1 over 60%)
- Although there are successful attempts of extracting MWEs from Wikipedia, there is no effort allowing extraction of all MWEs from plain text.



Using the Wikipedia as a dictionary

- Filter out the headwords without semantic labels.
- Create patterns containing one entry for each token.
- For words that can be inflectable, the pattern contains the grammatical categories.

For the headword "Związek Lekkoatletyczny Krajów Bałkańskich" the pattern is shown below in a simplified form. '?' means *maybe inflectable*. It is not possible to tell if "związek" should be capitalised because all Wikipedia headwords are capitalised.

Example

[? **(z|Z)wiązek** *male non-animate singular noun*]
[? **Lekkoatletyczny** *male singular adj.*] [**Krajów**] [**Bałkańskich**]

Dictionary pattern matching (DM)

- A non-deterministic finite automaton with one transition per pattern option for each token
- Fixed tokens have to be the same as the pattern entries
- Inflectable words have to be inflected correctly

Positive example

"Związkiem Lekkoatletycznym Krajów Bałkańskich" (instrumental case)

Negative example

"Związkiem Lekkoatletyczną Krajów Bałkańskich" (second word changed gender to feminine)

Negative example

"Związkiem Lekkoatletycznemu Krajów Bałkańskich" (first word in the instrumental case, second in the accusative case)

Test on PAP corpus

- No known corpora of Polish text tagged for MWEs, so we needed to create the test set from a corpus.
- Testing on Wikipedia was not easy as it cutting out the test sample would break its network structure.
- We decided to use a corpus of press releases of the Polish Press Agency (PAP).
- A random sample of 100 releases was tagged by two annotators.

An example of a tagged press release

We wtorek **[*mistrz *olimpijski]**, świata i Europy w chodzie **[*Robert *Korzeniowski]** weźmie udział w stołecznym **[*Lesie *Kabackim]** w **[*Biegu ZPC SAN]** o **[*Grand *Prix]** Warszawy oraz o **[*Grand *Prix] [*Polskiego *Związku Lekkiej Atletyki]**.

Syntactic pattern matching (SM)

- DM is limited to the headwords of Wikipedia: $P = 84.6\%, R = 37.1\%, F1 = 51.7\%$
- Groups of MWEs often share similar syntactic structure: "tlenek węgla" (*carbon dioxide*) and "chlorek sodu" (*sodium chloride*)
- We decided to automatically create **syntactic patterns** that would express these regularities.
- The matching would work similarly to DM

Example

The pattern for "czarna dziura" (*black hole*) is:

[fem. singular adj.] [* fem. singular noun]*

Context patterns

- Groups of MWEs often appear in text in similar context – e.g. in genitive case after a noun ("roztwór chlorku sodu" – *a solution of sodium chloride*, "reakcja tlenku węgla" – *a reaction of carbon dioxide*)
- We could extend the concept of syntactic patterns by including the surrounding words for the appearances of the MWE
- The context covers one token before and one after the MWE

Example

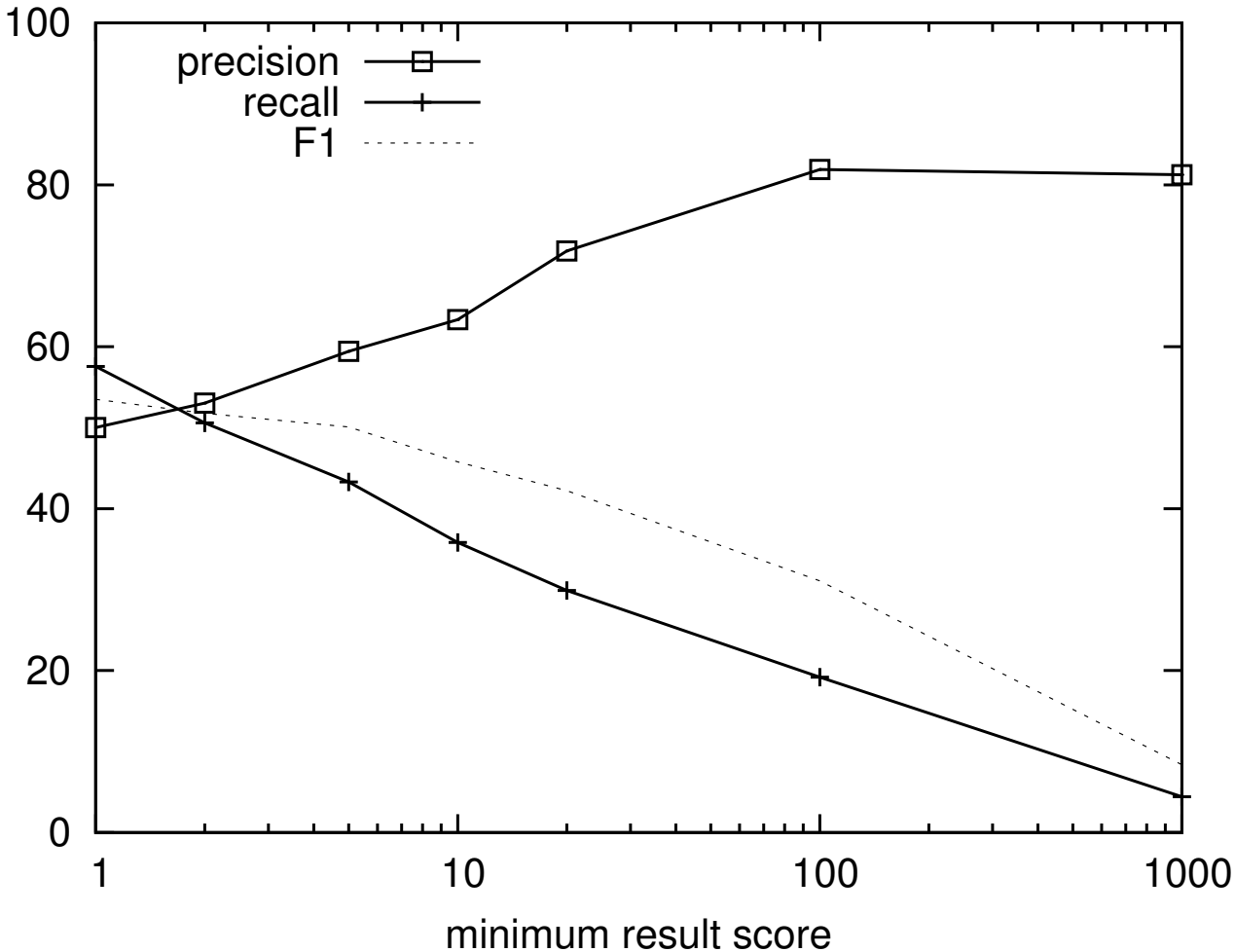
If "czarna dziura" occurs in the expression: "(...) masywnej **czarnej dziury**.", the context pattern will be:

[fem. singular noun, gen., acc. or loc.] ***[* fem. singular adj.] [* fem. singular noun]****[punct.]*

Syntactic pattern construction

- We need occurrences of Wikipedia headwords in text.
- We chose to look for this information in the **incoming links** and occurrences of the headword in the article content.
- The inflected occurrences are used to disambiguate the pattern.
- The algorithm is complex and has to find the largest subset of occurrences that lead to a non-contradictory pattern.
- The headwords used to create patterns are filtered using multiple thresholds, leaving only about 150 thousand out of original 550 thousand entries.
- For each matched word the SM automaton can calculate a **score**
  - total number of all occurrences of the matched patterns for the recognised form

Syntactic pattern matching (SM) results



Dictionary matching alternative: pDM

pDM

Creating the syntactic patterns leaves only the high quality entries, so why not use only them for DM?

- DM for words with syntactic patterns (**pDM**)
- + Allows identifying inflectable tokens and lowercase first tokens
- + Allows scoring the results according to the number of occurrences of the pattern for the given word
- − Less words: low recall ($P = 90.72, R = 29.88, F1 = 44.96$)

Improving syntactic patterns: SM-SM

Idea: create more patterns from the words matching the ones we already have.

SM-SM

Run SM on a corpus of all Wikipedia articles (WC) to obtain a dictionary of all matching words. Then create additional syntactic patterns from that dictionary.

- $+$ More pattern occurrences $\Longrightarrow$ better scores for patterns
- $-$ More complicated
- $-$ Additional step lowers precision

Improving syntactic patterns: pDM-SM

Idea: create patterns from occurrences of the headwords in **all** Wikipedia articles.

pDM-SM

Run pDM on a corpus of all Wikipedia articles (WC) to obtain a dictionary of all matching words. Then create additional syntactic patterns from that dictionary.

- + More occurrences $\Longrightarrow$ richer data, so higher recall
- − More complicated
- − Additional step and inclusion of lower quality occurrences lowers precision

Improving syntactic patterns: DM-SM

Idea: create patterns from occurrences of **all** Wikipedia headwords in all Wikipedia articles.

DM-SM

Run DM on a corpus of all Wikipedia articles (WC) to obtain a dictionary of all matching words. Then create additional syntactic patterns from that dictionary.

- $+$ More headwords and more occurrences $\Longrightarrow$ even higher recall
- $-$ More complicated
- $-$ Additional step and inclusion of lower quality entries lowers precision

Using syntactic patterns to improve the dictionary: SM-DM

Idea: extend the dictionary with all words matching the syntactic patterns.

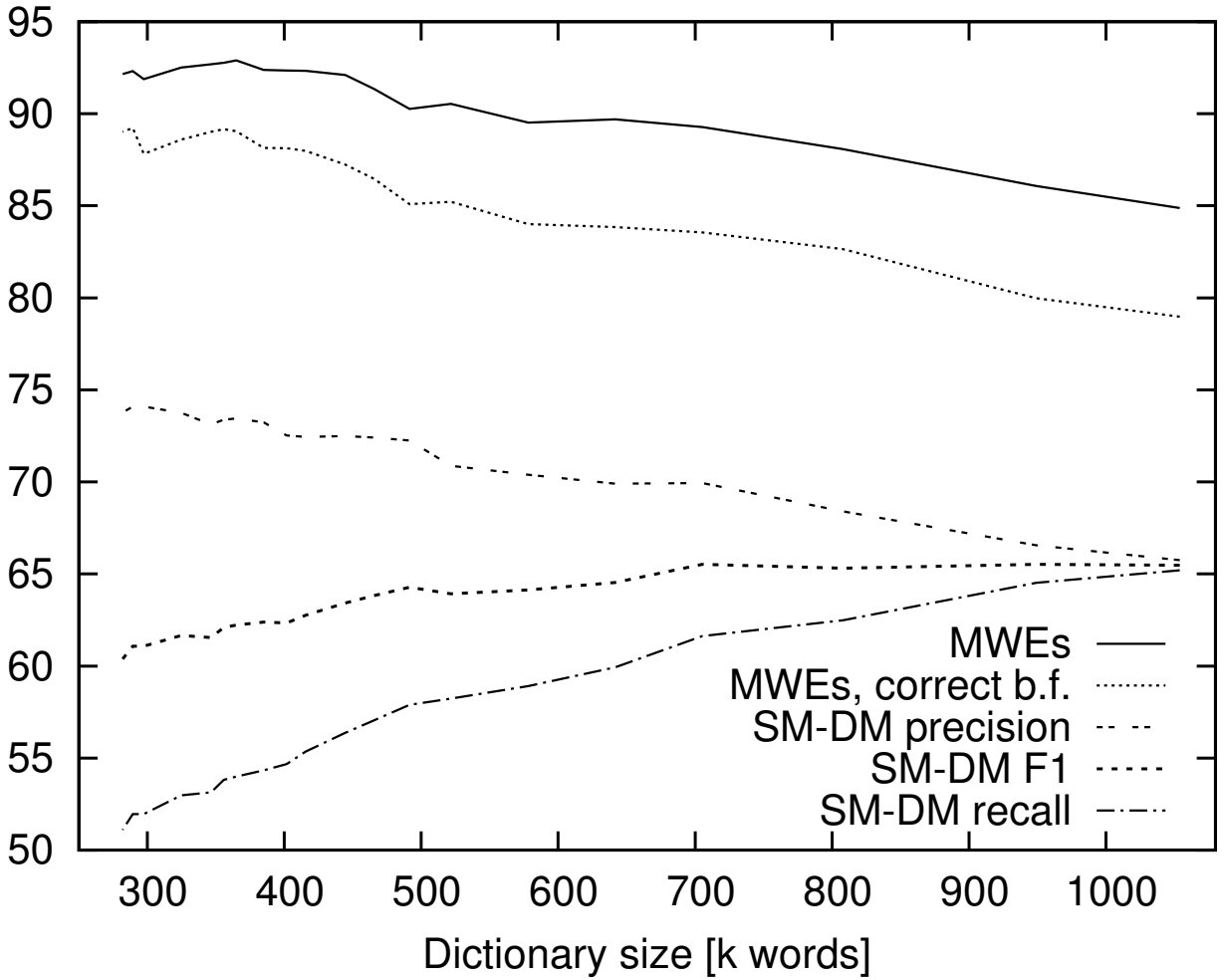
SM-DM

Run SM on a corpus of all Wikipedia articles (WC) to obtain a dictionary of all matching words. Then use these words to extend the pDM/DM methods.

- $+$ Much more words than for pDM or DM
- $-$ Additional step lowers the precision

SM-DM results

- SM phase produces a good quality MWE dictionary
- DM benefits from using that dictionary



Test result comparison

- **Structural test** – only the tokens of the MWE need to be correct
- **Syntactic test** – the inflectable words have to be correctly identified

	Structural test			Syntactic test		
Algorithm	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
SM-DM	65.60	66.38	65.99	56.21	56.88	56.54
pDM-SM	48.99	61.97	54.72	41.61	52.63	46.48
SM-SM	52.99	55.69	54.30	44.59	46.86	45.70
DM-SM	48.97	60.44	54.10	40.85	50.42	45.14
SM	50.00	57.56	53.51	42.18	48.56	45.15
SM0	49.85	56.54	52.98	41.32	46.86	43.91
DM	84.56	37.18	51.65	56.76	24.96	34.67
pDM	90.72	29.88	44.96	86.08	28.35	42.66

Test result analysis

Why is F1 so low (66%)?

- **Overlapping expressions** difficult to choose from (even for a human): "Piłkarska reprezentacja Brazylii" contains two MWEs: "piłkarska reprezentacja" (*a football team*) and "reprezentacja Brazylii" (*Brazilian team*).
- Unusually **long names**, **spelling and structural errors**: "Doroty Kędzierzwskiej" (misspelled surname), "Polska Fundacja Pomocy humanitarnej "Res Humanae"" (unusual structure and an accidental lowercase letter).
- Phrases that are not MWEs in the particular (semantic) context: "osoba paląca sporadycznie" contains the term "osoba paląca" (*a smoker*) while it means *a person smoking sporadically*.
- SFJP is missing proper names (mostly surnames): "Janusz Steinhoff", "Władysław Bartoszewski".

Future work and improvement

- This is exploratory work – not quite ready for real world applications
- There is a precision-recall tradeoff: supervised methods such as SVM could provide better classification than the simple threshold used now
- SFJP does not contain enough proper names – it should be extended or other tools (eg. Morfologik) could be used

Conclusions

- Inflection dictionaries are useful for natural language processing, but they lack **multiword expressions** (MWEs).
- We used the Wikipedia to create extract MWEs from Polish text:
  1. **Syntactic patterns** can be created from selected Wikipedia headwords
  2. The patterns can be recognized in the whole Wikipedia content to create an inflection dictionary of MWEs
  3. This dictionary can be used to extract MWEs from Polish text with F1=66%

References I

Attia, M., Tounsi, L., Pecina, P., van Genabith, J., and Toral, A. (2010)
Automatic extraction of arabic multiword expressions.
In: Proceedings of the Workshop on Multiword Expressions: from Theory to Applications, Beijing, China, 28 August 2010, pp. 18–26.

Chrząszcz, P. (2012)
Enrichment of inflection dictionaries: automatic extraction of semantic labels from encyclopedic definitions.
In: Sharp, B. and Zock, M. (eds) Proceedings of the 9th International Workshop on Natural Language processing and Cognitive Science, Wrocław, Poland, 28 June – 1st July 2012, pp. 106–119.

Church, K. (2013)
How many multiword expressions do people know?
ACM Transactions on Speech and Language Processing (TSLP), 10(2):4.

References II

Constant, M., Sigogne, A., and Watrin, P. (2012)
Discriminative strategies to integrate multiword expression recognition and parsing.
In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers – Volume 1, Jeju, Korea, 8–14 July 2012, pp. 204–212.

Farahmand, M. and Martins, R. (2014)
In: A supervised model for extraction of multiword expressions based on statistical context features.
Proceedings of the 10th Workshop on Multiword Expressions, Gothenburg, Sweden, 26–27 April 2014, pp. 10–16.

Gajęcki, M. (2009)
Słownik fleksyjny jako biblioteka języka c.
In: Lubaszewski, W. (ed) Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu, pp. 107–134. Wydawnictwa AGH, Kraków.

References III


Kuta, M., Chrząszcz, P., and Kitowski, J. (2007)
A case study of algorithms for morphosyntactic tagging of polish language.
Computing and Informatics, 26(6):627–647.


Lubaszewski, W. (2009)
Wyraz.
In: Lubaszewski, W. (ed) Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu, pp. 15–36. Wydawnictwa AGH, Kraków.


Lubaszewski, W., Wróbel, H., Gajęcki, M., Moskal, B., Orzechowska, A., Pietras, P., Pisarek, P., and Rokicka, T. (2001)
(ed) Słownik Fleksyjny Języka Polskiego.
Lexis Nexis, Kraków.


Nothman, J., Murphy, T., and Curran, J. R. (2009)
Analysing wikipedia and gold-standard corpora for ner training.
In: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, Athens, Greece, 30 March – 3 April 2009, pp. 612–620.

References IV

Piskorski, J., Homola, P., Marciniak, M., Mykowiecka, A., Przepiórkowski, A., and Woliński, M. (2004)
Information extraction for polish using the sprout platform.
In: Kłopotek, A., Wierzchoń, T. and Trojanowski, K. (eds) Proceedings of the International IIS: IIPWM' 05 Conference, Gdańsk, Poland, 13–16 June 2005, pp. 227–236.

Pohl, A. (2009)
Rozstrzyganie wieloznaczności, maszynowa reprezentacja znaczenia wyrazu i ekstrakcja znaczeń
In: Lubaszewski, W. (ed) Słowniki komputerowe i automatyczna ekstrakcja informacji z tekstu, pp. 241–255. Wydawnictwa AGH, Kraków.

Ramisch, C., Schreiner, P., Idiart, M., and Villavicencio, A. (2008)
An evaluation of methods for the extraction of multiword expressions.
In: Proceedings of the LREC Workshop – Towards a Shared Task for Multiword Expressions, Marrakech, Morocco, 1 June 2008, pp. 50–53.

References V

Richman, A. E. and Schone, P. (2008)
Mining wiki resources for multilingual named entity recognition.
In: Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Columbus, OH, USA, 15–20 June 2008, pp. 1–9.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002)
Multiword expressions: A pain in the neck for nlp.
In: Gelbukh, A. (ed) Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, 17–23 February 2002, pp. 1–15.

Tjong Kim Sang, E. F. and De Meulder, F. (2003).
Introduction to the conll-2003 shared task: Language-independent named entity recognition.
In: Daelemans, W. and Osborne, M. (eds) Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 27 May – 1 June 2003, Edmonton, Canada, Vol. 4, pp. 142–147.

References VI

Woliński, M. (2006)
Morfeusz – a practical tool for the morphological analysis of polish.
Advances in Soft Computing, 26(6):503–512.

Zhang, Y., Kordoni, V., Villavicencio, A., and Idiart, M. (2006)
Automated multiword expression prediction for grammar engineering.
In: Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties, July 2006, Sydney, Australia, pp. 36–44.

Zhou, G. and Su, J. (2002)
Named entity recognition using an hmm-based chunk tagger.
In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, 7–12 July 2002, Philadelphia, PA, USA, pp. 473–480.