Discovery of Common Nominal Facts for Coreference Resolution: Proof of concept[*]

Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences

Abstract. This paper reports on the preliminary experiment aimed at verification whether extraction of nominal facts corresponding to world knowledge from both structured and unstructured data could be effectively performed and its results used as a source of pragmatic knowledge for coreference resolution in Polish. Being the proof-of-concept only, this approach is work in progress and is intended to be further validated in a full-scale project.

1 Introduction

Coreference resolution is traditionally defined as a process of determining which fragments of a text correspond to the same discourse-world entities. As such, it is usually performed in two steps:

1. identifying *mentions* (or *markables*), i.e. phrases denoting entities in question
2. clustering mentions which denote the same referent.

The current scope of interest in research on coreference resolution for Polish is direct nominal coreference, i.e. identity-of-reference (in contrast to other anaphoric phenomena such as identity-of-sense anaphora, ellipsis, bound anaphora or bridging anaphora), with mentions being nominal groups (including single nouns, pronouns etc.). Following this assumption, the Polish Coreference Corpus [1] and coreference resolution tools [2,3] have been created, offering possibility to continue research on the subject.

The state-of-the art coreference resolution tools for Polish employ four extensive groups of features:

1. surface features (e.g. linking orthographic entity name with its abbreviation)
2. syntactic features (e.g. traditional gender/number agreement)
3. semantic features (e.g. agreement between semantic classes of mention heads)
4. discourse features (e.g. salience of topics).

Such approach results in a sufficiently effective (as compared to other languages) resolution process, but analysis of remaining errors reveals its one shortage: lack of

[*] The work reported here was carried out within the *Computer-based methods for coreference resolution in Polish texts (CORE)* project financed by the Polish National Science Centre (contract number 6505/B/T02/2011/40).

representation of the world knowledge leads to clustering misses, affecting the final score of the whole process. Introducing pragmatic features representing widely known facts would, as we believe, increase probability of linking mentions denoting the same discourse-world object. In this paper we intend to verify whether this assumption is true before it can be applied in a large-scale project.

2 Analysis of the Problem: Going Beyond Semantics

Currently available semantic bases such as WordNet (and its Polish equivalents: plWordNet [4] and POLNET [5]) only partially resolve this issue, by offering coarse semantic classes and semantic network traversal with via hypo-/hyperonymy/synonymy relations, which is not sufficient for texts abundantly using semantically close nominal phrases, making their automatic clustering problematic.

At the same time, it very often happens that there is no semantic connection between coreferent phrases whatsoever. In the following example:

(1) *Aldrin and Armstrong przyjaźnili się nadal, mimo że cała uwaga mediów skupiła się wyłącznie na pierwszym człowieku na Księżycu.*

'Aldrin and Armstrong stayed friends even though the whole attention of media now focused on the first man on the Moon.'

the resolution going beyond a random guess is not possible when only features from the four above-mentioned groups are applied. It is nevertheless true that linking *Armstrong* and *the first man on the Moon* could be easy for most human annotators — and even some search engine-based systems, even using the simplest full-text search mode.

Sometimes the situation gets complicated by the nature of the domain; the phrases *Adam Mickiewicz*, *the husband of Celina Szymanowska*, *the poet*, *the lecturer in College of France* can be clustered together only with some (deeper) knowledge about the life of Adam Mickiewicz, a Polish 19th century poet. Without referencing the history of Polish literature both a person and a computer system would experience difficulties to resolve coreference between those phrases. However, the border between common and specific knowledge is vague, especially in the face of availability of such resources as Wikipedia, offering ready-to-extract information on even less-generally known topics.

We deliberately skip one more (rare) case which should be noted for completeness: understanding of certain concepts in expert knowledge can be different from 'the common knowledge', which may hinder coreference resolution. For example, in the scientific sense tomato is the fruit (mature ovary) of the tomato plant, but in common interpretation (e.g. in cooking) tomato is a vegetable. We leave these difficult cases aside as they go beyond the scope of this paper.

3 Concept of the Pragmatic Nominal Knowledge Base

Since the nature of coreference resolution problem is conceptual: establishing and decoding coreference is about sharing the same knowledge of discourse entities between the speaker (conveying some message in the text, being the primary communication channel) and the recipient (decoding method), we could make an attempt at establishing a common, reusable, updateable platform of understanding of the facts expressed in the text being analysed. Within the scope of the resolution task, limited to nominal groups, such platform could be conceived of as a pragmatic knowledge base composed of 'seed' nominal facts and their interpretations.

This type of information goes far beyond semantic relations present e.g. in the Wordnet, with its Polish version unable to maintain definitions such as *pediatria* ('pediatrics') — *nauka o chorobach dziecięcych* ('branch of medicine that deals with child's diseases'). Similarly, this information cannot be inferred from investigating syntactic heads of phrases since *człowiek* ('man') carries much different information capacity than the whole phrase *pierwszy człowiek na Księżycu* 'the first man on the Moon'.

The content of such base would cover established facts (such as, again, linking Neil Armstrong with his well-known attribute of being "the first man on the Moon") and typical periphrastical realisations of frequent nominal phrases, including named entities (e.g. linking Napoleon Bonaparte with his nickname, "The Little Corporal").

3.1 Data Extraction Sources

To boost development of the knowledge base, we plan to reuse existing sources of structured and unstructured data which now has been used for years in construction of semantic lexicons [6], information extraction [7] or Web question answering [8].

Structured sources should be represented by existing data- and knowledge repositories such as traditional dictionaries. For Polish two adequate resources of these type are: The Dictionary of Periphrastic Constructions [9] and The Great Dictionary of Polish — WSJP [10], both prepared by the scientific community. On the other hand, there is a growing number of crowd-sourced dictionaries and definition bases, in most cases intended to be used for Internet games and crosswords (`http://sjp.pl`, `http://krzyzowki.info`). Processing data from these groups would consist in automatic filtering of nominal definitions and passing them to manual verification.

Digital ontologies (explicit specifications of conceptualization) could also be used as source of periphrases, most likely with typical nominal instantiations of knowledge items generated in a human-readable form (to be later matched with textual content), but we deliberately omit this method, on one hand because of mostly derivative nature of such resources and on the other — due to their artificial character, abstracted from realistic use of language. To illustrate this problem, let's analyse the relation between the phrase *gród Kraka* ('Krak's (fortified) town') and its synonym, city name *Kraków* ('Cracow'). The former one is frequently used in texts about Cracow to maintain cohesion but we could hardly ever find it when looking in available structured sources. Moreover, it will never be automatically generated from any ontology because of its collocational character and atypical component *gród*, rarely used in a contemporary texts when referring to a town.

Capturing phenomena of this type can only be achieved by processing unstructured sources representing the bottom-up approach to language and likely to enrich dictionary data with real-life examples. In the long run we plan to process both balanced corpora (such as NKJP [11] or KPWr [12], providing standard representation of a language), available content sources (such as Gutenberg project) and sources of dynamic language — electronic media archives such as *Korpus Rzeczpospolitej* [13] or current parliamentary transcripts from the Polish Sejm Corpus [14].

4 The Experiment

Our hypothesis was that pragmatic data available in online data sources could improve coreference resolution in Polish by providing associations unavailable to obtain with currently used methods (surface, syntactic, semantic or discourse-based). To verify that, we have compared manual annotation of nominal mentions in the corpus of general nominal coreference — Polish Coreference Corpus, PCC [15] with their automatic annotation created with Ruler [2] to extract coreferential links identified by human annotators, but missed by the computer resolver and having the property of semantic unrelatedness. Absence of such link gives sufficient indication that the current resolution methods could not create the association, but there exists some additional level of understanding of the text which makes it obvious for the human annotators.

Out of 1220 nominal clusters (with only nominal mentions) 73 mention pairs have been manually selected for further processing. They constituted all data for which coreference resolution was unfeasible with the above-mentioned means. Mentions which are currently not clustered, but could get resolved with additional semantic effort, were removed from the data set. Two examples of such semantic-intensive data are *czternaście tysięcy złotych* ('fourteen thousand Polish zlotys') — *pierwsza tak duża dotacja* ('the first so huge subsidy'), when cluster could have been created by comparing wordnet-based semantic classes, and *marszałek* ('marshal') — *Marek Nawara, marszałek małopolski* ('Marek Nawara, the marshal of Małopolska')[1], when appositional components could have been inspected to create the link.

4.1 Data Classification

It occurred that the contents of the set follows, to a great extent, the common classifications of named entities, such as the one used for the National Corpus of Polish (see e.g. [16]). Among the 73 mention pairs included in the set, four out of five following subclasses are named entity-related and follow the NKJP classification:

- 29 personal names linked with person role, function, occupation etc. (e.g. *Jan Paweł II* ('John Paul II') — *polski papież* ('the Polish pope'), *Rafał Blechacz* — *pianista ogromnie utalentowany i skromny* ('a pianist tremendously talented and modest')
- 18 names of organisations — companies, sports clubs, political parties, music bands etc. (e.g. *Ich Troje* — *zespół Michała Wiśniewskiego* ('Michał Wiśniewski's band'), *Wizzair* — *tania linia lotnicza* ('low-cost airline'))
- 14 geographical/geo-political names — here: only names of countries and cities (e.g. *Irak* ('Iraq') — *kraj* ('country'), *Aleksandrów Łódzki* — *miasto* ('city'))

[1] In PCC appositions are treated as components of the main phrase.

- 6 'human creation' names — movie, book and newspaper titles (e.g. *Star Trek* — *dzieło filmowe* ('cinematographic work'), *Wahadło Foucaulta* ('Foucault's Pendulum') — *książka* ('a book'))
- 6 descriptive definitions, e.g. *kot* ('cat') — *udomowiony ssak* ('domesticated mammal')), *lekarze i pielęgniarki* ('doctors and nurses') — *personel szpitalny* ('hospital staff')).

Such statistics imply that the seed concepts should be closely related to named entities. It results in the first place from absence or underrepresentation of such concepts in the Polish WordNet — quoting city examples, plWordNet contains 339 sample instances of the artificial synset *miasto Polskie* ('a Polish city') which corresponds to 1/3 of the total number of all cities in Poland. Nevertheless, the structure and contents of any wordnet cannot be subordinated to ideology of representing the whole world knowledge – c.f. the Princeton WordNet, similarly far from representing company names or movie titles.

4.2 Knowledge Extraction Attempt

Each of the mention pairs have been manually tested against one of the knowledge bases mentioned in Section 3.1 to provide a proof of concept that extracted data used as 'pragmatic features' would, to a large extent, help in proper clustering of mentions in the coreference resolution process.

2 sources have been selected as main supplies of pragmatic data: Polish Wikipedia and online crossword definition service `http://krzyzowki.info`. This decision was based on the assumption that Wikipedia is a reliably enough source of information about named entities while crossword services should provide sufficient support for definitions. Table 1 provides statistics of data sources used for resolving mention pair dependency, showing number of entity pairs which could be resolved using only Wikipedia, only the crossword definitions, with both methods, some other algorithmically available method or which could not be resolved by any pragmatic means.

Table 1. Sources of pragmatic information

	Wikipedia	**krzyzowki.info**	**both**	**other**	**none**
personal names	14		14	1	
organisations	9		8		1
geo names		1	13		
creation names	1		5		
definitions	4		1	1	

The first important finding is that all but one problematic assignments could be properly resolved; the missing one resulted from manual annotation error (wrong association between a soccer club name and a mountain name: *Klimczok*). Another striking fact is that for most mention pairs (all but three) the resolution process could be completed by using only Wikipedia. The only definition-based case was the association between the country name *Niemcy* ('Germany') and its property: *zachodni sąsiad Polski* ('the western neighbour of Poland'), possible to get resolved using the textual head-match with the phrase present in the definition base.

'Other' resolution source indicates that both of the main sources were not sufficient to resolve the link, but another available online source could be used; the examples here are diminutive and augmentative form of the name *Małgorzata* ('Margaret'): *Gosia* and *Gocha* and a common name for medical staff: *lekarze i pielęgniarki* ('doctors and nurses') — *personel szpitalny* ('hospital staff').

4.3 Data Abstraction

When collected, separate set of algorithms can be used to abstract nominal facts from nominal phrases which we believe to boost coreference resolution recall while maintaining storage efficiency. Apart from typical collocations which should only be processed in a controlled manner, two abstraction components are now envisaged: a syntactic one and semantic one.

The former would convert between different syntax models of a phrase maintaining its meaning, e.g. relative to participial phrases: *osoba, która podpowiada aktorom* ('a person who feeds lines to actors') — *osoba podpowiadająca aktorom* ('a person feeding lines to actors'). The semantic component would use wordnet relations such as synonymy or hyponymy to neutralise lexical meanings of phrase components: *osoba, która podpowiada wykonawcom* ('a person who feeds lines to performers').

Evaluation of both components would be a starting point for further investigation of several independent research problems e.g.:

- how alternation of verbal constructs influences usage of phraseology (*przejąć* ('to take over') — *dokonać przejęcia* ('to make a takeover'), *człowiek, który przepłynął Atlantyk* ('a man who sailed across the Atlantic') — *człowiek, który przebył Atlantyk* ('a man who travelled across the Atlantic'))
- how far can attributes modify nominal syntax constructs (*mała niebieska pigułka* ('little blue pill') — *niebieska pigułka* ('blue pill')
- which factors influence syntactic stability of collocations (cf. *Kraj Wschodzącego Słońca* ('Land of the Rising Sun')).

5 Conclusions and Further Work

The experiments confirmed our original hypothesis that currently available data sources can provide pragmatic knowledge and in this way improve coreference resolution in Polish when currently used algorithms fail. Apart from coreference resolution, the completed version of the database will also find its other linguistic applications such as pragmatic analysis of text for smoothing the result of automatic text summarization, machine translation or readability improvements. Development of the knowledge base would seriously enrich capabilities of independent IT systems performing text analysis, especially as current version of such systems are insensitive to pragmatic facts vital for correct interpretation of the text while such information is freely available to all search engines, even in the simplest full-text search mode (cf. search results for „Nazi Propaganda Minister"). The knowledge base could also be made available independently, in a WolframAlpha-like interface offering search and visualisation.

Considering incremental and volatile character of knowledge, expressed by constant update of underlying resources by Internet users, extraction algorithms could be linked with data sources in a way triggering updates of the knowledge base contents when source data (e.g. Wikipedia article) gets updated.

The data pool could be extended with more linked data sets and tools traditionally used for ontological modelling, with possibility of using ontological relations to improve data abstraction (e.g. when *Księżyc* ('the Moon') is linked in ontology to *Srebrny Glob* ('the Silver Globe'), it could be used in abstraction of phrases like *pierwszy człowiek na Księżycu* ('the first man on the Moon')). Interfacing with WolframAlpha or Google Knowledge Graph will be also investigated. Last but not least, foreign-language resources could be examined to import translated nominal representation of knowledge bits to the base.

References

1. Ogrodniczuk, M., Zawisławska, M., Savary, A., Głowińska, K.: Coreference Annotation Schema for an Inflectional Language. In Gebulkh, A., ed.: Proceedings of the 14th International Conference on Intelligent Text Processing and Computational Linguistics. Part I. Volume 7816 of Lecture Notes in Computer Science., Heidelberg, Springer-Verlag (2013) 394–407
2. Ogrodniczuk, M., Kopeć, M.: End-to-end coreference resolution baseline system for Polish. In Vetulani, Z., ed.: Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, Poznań, Poland (2011) 167–171
3. Kopeć, M., Ogrodniczuk, M.: Creating a Coreference Resolution System for Polish. [17] 192–195
4. Piasecki, M., Szpakowicz, S., Broda, B.: A Wordnet from the Ground Up. Oficyna Wydawnicza Politechniki Wrocławskiej (2009) `http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf`.
5. Vetulani, Z., Kubis, M., Obrębski, T.: PolNet — Polish WordNet: Data and Tools. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D., eds.: LREC, European Language Resources Association (2010)
6. Thelen, M., Riloff, E.: A bootstrapping method for learning semantic lexicons using extraction pattern contexts. In: Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10. EMNLP '02, Stroudsburg, PA, USA, Association for Computational Linguistics (2002) 214–221
7. Agichtein, E., Gravano, L.: Snowball: extracting relations from large plain-text collections. In: Proceedings of the fifth ACM conference on Digital libraries. DL '00, New York, NY, USA, ACM (2000) 85–94
8. Dumais, S., Banko, M., Brill, E., Lin, J., Ng, A.: Web question answering: is more always better? In: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '02, New York, NY, USA, ACM (2002) 291–298
9. Bańko, M.: Słownik peryfraz czyli wyrażeń omownych. PWN Scientific Publishers, Warszawa (2003)

10. Żmigrodzki P.: O projekcie Wielkiego słownika języka polskiego. Język Polski **5**(LXXXVII) (2007) 265—-267
11. Przepiórkowski, A., Bańko, M., Górski, R.L., Lewandowska-Tomaszczyk, B., eds.: Narodowy Korpus Języka Polskiego [Eng.: National Corpus of Polish]. Wydawnictwo Naukowe PWN, Warsaw (2012)
12. Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., Wardyński, A.: KPWr: Towards a Free Corpus of Polish. [17] 3218–3222
13. Presspublica: Korpus Rzeczpospolitej. [on-line] `http://www.cs.put.poznan.pl/dweiss/rzeczpospolita`
14. Ogrodniczuk, M.: The Polish Sejm Corpus. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA) (2012)
15. Ogrodniczuk, M., Głowińska, K., Kopeć, M., Savary, A., Zawisławska, M.: Interesting Linguistic Features in Coreference Annotation of an Inflectional Language. In et al., M.S., ed.: CCL and NLP-NABD 2013. Volume 8202 of Lecture Notes in Computer Science. Springer-Verlag, Berlin, Heidelberg (2013) 97–108
16. Waszczuk, J., Głowińska, K., Savary, A., Przepiórkowski, A., Lenart, M.: Annotation tools for syntax and named entities in the National Corpus of Polish. International Journal of Data Mining, Modelling and Management **5**(2) (2013) 103–122
17. Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., eds.: Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012, Istanbul, Turkey, ELRA, ELRA (2012)