

Proper Names in Dialogs from the Warsaw Transportation Call Center

Małgorzata Marciniak, Joanna Rabeiga-Wiśniewska,
and Agnieszka Mykowiecka

Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland

Abstract

In the paper we present the method of automatic recognition and annotation of proper names which occur in dialogs gathered at the Warsaw city transportation information center. We describe different types of proper names and how people use them in dialogs. We present rules of automatic recognition and lemmatization of proper names in the transportation domain.

Keywords: proper names, dialog corpora annotation

1 Introduction

The paper concerns recognition and annotation of proper names that occur in Polish real-life dialogs concerning city transportation in Warsaw.¹ The dialogs were collected at the information call center where people can get information on: routes between the given points of the city; time schedules of public transport; an expected trip duration; stops; tickets and reduced-fares. The corpus (Mykowiecka *et al.*, 2007) consists of directories containing the following files for each dialog:

- an acoustic signal of the dialog;
- transliteration and annotation on acoustic level of the dialog;
- information about turns boundaries and number of words adjusted to each of them;
- information on POS of each word, its base form and morphological characteristics;
- elementary syntactic chunks e. g.: elementary nominal, numeral, verbal, adjectival phrases, and proper names;
- annotation on the level of the domain attributes (concepts) that consists in assigning attributes' names from the transportation ontology to phrases from the dialog.

City transport is an area in which usage of proper names is very frequent. Although their recognition can be simplified by using appropriate databases, in this particular case all typical NER problems can be observed. In the dialogs,

¹This work is supported by the LUNA project (IST 33549, www.ist-luna.eu).

new names appear infrequently, but a lot of names are used in short unofficial forms. Some proper names can occur within other names and the same strings can refer to different concepts, e. g.: *Tor Stegny* is the name of the building or the name of the bus stop and *Stegny* is the name of the district. Additionally, Polish names undergo inflection which makes their recognition even more challenging. To connect various forms to the appropriate objects it is necessary to change inflected forms into canonical ones.

The paper describes the process of domain attributes annotation of the dialogue corpus limited to the problem of proper names. We describe different types of named entities and their variants used by people in spontaneous speech. Then, we present our automatic method of NER based on manually created rules. There are three types of such rules: rules that recognize proper names without any additional information, rules that use introductory words such as *ulica* (street), *kierunek* (direction), and rules that use some context information like *przystanek nazywa się Uniwersytet* (the stop name is Uniwersytet). The final part of the paper describes lemmatization of proper names.

2 NE Types and Recognition Problems

In the selected domain the most frequent name entity types are:²

- names of streets, e. g. *Marszałkowska, Bitwy Warszawskiej 1920 r., Zygmunta Krasińskiego*;
- names of towns and city districts, eg. *Wesoła, Stara Miłosna, Wola, Dolny Mokotów*;
- names of buildings, e. g. *Teatr Komedia, Galeria Mokotów, CH Wola Park*;
- names of transportation lines, e. g. linia 501, E-2, piętnastka.

Below, we present main problems with recognizing NEs in the selected type of dialogs:

- a name structure can be complicated and two names can occur in a sequence, e. g. *linia E-1 Rondo ONZ*;
- one name can be used in different inflectional forms, e. g. *Wesoła, Wesołej, Wesołą* or variants, e. g. *Żwirki i Wigury* vs. *Żwirki*;
- one name can contain another one:
 - names of streets or buildings can contain names of persons or organizations, e. g. *ulica Zygmunta Krasińskiego, Plac Krasińskich, Rondo ONZ*;
 - names of stops are mostly based on names of buildings, streets, metro stations etc. *Metro Imielin, przystanek Metro Imielin, przystanek Metro Imielin zero dwa*;
- in street names containing person names, first names are nearly always omitted in speech, e. g. *Adama Mickiewicza* → *Mickiewicza*, but for some streets first names are used quite often, e. g. *Emilii Plater*.

²Description of streets and districts names in Warsaw is given in Handke 1998, see also Rabiega-Wiśniewska 2008 for a review of proper names contained in the dialogs.

TABLE 1: Size of lexicons

type	total	one word	multi words
street names	3983	3883	100
towns and city districts	145	113	32
building names	147	48	99
area names	14	1	13

To be able to recognize all these proper names we: created lists of names of places in the city; enriched the morphological lexicon with proper names' forms; and defined contextual rules which recognize different ways of addressing particular names. These rules assign domain dependent concepts to the recognized word groups. Among these concepts there are 7 types of proper names: \$STREET, \$TOWN, \$TOWN_DISTRICT, \$BUILDING, \$AREA (other place names), \$TRANSPORTATION_LINE and \$STOP_DESC. Numbers of some NE classes are presented in Table 1.

3 Annotation Rules

Proper name recognition is performed during a rule based semantic annotation of texts with concept names. In this section we describe the syntax of annotation rules. They operate on disambiguated morphological tags and chunk boundaries (and indirectly on turn boundaries). As they do not allow recursion, three sets of rules have been defined and they are applied in subsequent passes of the annotation process. Before their application, the described above lexicons are checked and appropriate concepts' names are assigned to the forms found in a dialog.

Each annotation rule consists of three parts: (a) a concept name that will be assigned to a recognized phrase, (b) its value that can be constructed of several elements joined by '+', (c) and finally, after '#' symbol, the phrase elements.

(1) NAME output_value #pattern_description

The *output_value* can be a sequence of strings or variables which refer to elements of *pattern_description*. Numbered variables, \$1, \$2, \$3 etc., correspond respectively to subsequent pattern's elements (by default the lemma of the element is taken, if input form is needed it is described as \$n(form)).

A pattern describes conditions which have to be fulfilled by input elements. Optional elements are included in parenthesis and followed with '?'. Each pattern element can be:

- a variable – \$name to which a value was assigned by a rule from a lower level rule sets;
- a word with a fixed part of speech;
- a chunk type.

Pattern elements can include restrictions imposed on values of morphological features. Below we present three rules. The first two recognize street names and

the last one describes one of the methods of referring to a public transportation stop name.

- (2) \$STREET \$2 #(Nc(lemma=ulica))+\$STREET
 \$STREET \$2 #Np(lemma=Aleja,morph=pl)+ADJp
 \$STOP_DESC \$3 #Nc(lemma=przystanek)+\$STREET

The first rule above describes the possibility to add the word *ulica* (street) before the street name. It is useful in case when a plain name is ambiguous like *Wilanowska*. It can be the street: *ulica Wilanowska*, the avenue: *Aleja Wilanowska* and the metro station: *Metro Wilanowska*. The next one is an example of more relaxed rules and defines every sequence of a word *Aleja* and proper adjective as a street name. This rule is defined for recognizing streets which are not in our lexicon. The third rule describes a stop name based on a street name.

4 Proper Names Lemmatization

Lemmatization of proper names (assigning canonical forms), in Polish and other Slavic languages is a difficult problem because of reach inflection and relaxed word order. In English, the problem of proper names lemmatization is simple. Rule based lemmatization of French complex geographical names is described in Bellei and Maurel 1997. The systematic account of the syntax and semantics of names for different languages is given in Anderson 2007. A description of Polish proper names and their declensions is given in Grzenia 2003, and in Cieřlikowa 2008. Rather little work on this subject has been undertaken from the computational point of view. But some interesting results of experiments were presented in Piskorski et al. 2007 and Piskorski and Sydow 2007. The first paper describes a rule-based approach for lemmatization of Polish person names, and an approach based on different string distance measures inspired by the works of Cohen et al. (2003), and Christen (2006). The second paper presents an evaluation of different distance measures used for lemmatization. In the transport domain, most proper names concern location, and it turned out that their lemmatization differs from those for people's names.

In our data, the most problematic names which had to be lemmatized were street names (and stop names containing them). For most of them, assigning the nominative form is inappropriate, as many names of streets or stops are in the genitive form and they are not inflected, e. g. *Bonifacego* (not *Bonifacy*), or in plural, e. g. *Lotników* (not *Lotnik*). The canonical forms of adjectival names are mostly nominative but their gender depends on the nominal head. For example, the gender of *ulica* (street) is feminine, thus *ulica Lubelska*, but the gender of *trakt* (route) is masculine – so there is *Trakt Lubelski*. For other NEs (buildings, districts or line names) a nominative form is the most typical canonical form but these names are often composed of several words. In this case, the nominative form is almost always different from the sequence of nominative forms of its elements, eg. *z Sadów Żoliborskich, w Złotyach Tarasach* and *Bibliotekę Uniwersytetu Warszawskiego* should be lemmatized into *Sady Żoliborskie, Złote Tarasy* and *Biblioteka Uniwersytetu Warszawskiego* not into *Sad Żoliborski, Złoty Taras* and *Biblioteka Uniwersytet Warszawski*.

4.1 Dictionary Based Approach

In our approach, proper name recognition begins at the level of morphological annotation of the text. The inflectional lexicon (Rabiega-Wiśniewska and Rudolf, 2003) used for automatic morphological analysis of words has been extended with lexemes related to the domain of conversations. For proper names, new parts of speech have been introduced: a proper noun, number, preposition and adjective. The general rules for proper POS descriptions are the same as for common POS. For example, grammatical base forms for adjectives are in nominative masculine singular. The choice of this solution was caused by the fact that frequently different forms of one adjective appear in different proper names, e. g. a person name, *Krasiński*_{ADJ.nom.sg} is a part of a street name: ulica *Krasińskiego*_{ADJ.gen.sg}, a garden name: Ogród *Krasińskich*_{ADJ.gen.pl}, a square name: Plac *Krasińskich*_{ADJ.gen.pl} and a palace name: Pałac *Krasińskich*_{ADJ.gen.pl}. There are also feminine and masculine forms of adjectives that follow nouns with fixed gender, e. g. *ulica*_{fem} Lubelska vs. *Trakt*_{masc} Lubelski. What was convenient for minimizing the size of the lexicon, causes problems with the recognition of multi-word proper names. This is because there is no direct correspondence between a proper name of a place found in texts (*Aleja Dzieci*_{gen.pl} *Polskich*_{gen.pl.gender}, *Domy*_{pl} *Centrum*) and entry labels of their components gained from the lexicon (*Dziecko*_{nom.sg}, *Polski*_{nom.sg.masc}, *Dom*_{sg}).

Results of morphological annotation and disambiguation are further processed by the semantic annotation rules which were described briefly in the previous section. These rules operate on sequences of grammatical base forms and morphological information, see Table 2.

TABLE 2: Street names and their grammatical base forms

Street names	Base forms and inflectional values
Aleją Solidarności	Aleja:Np,nom.sg.fem Solidarność:Np,gen.sg.fem
Aleje Jerozolimskie	Aleja:Np,nom.pl.fem Jerozolimski:ADJp,nom.pl.nm1
Jana Sobieskiego	Jan:Np,gen.sg.m1 Sobieski:ADJp,gen.sg.masc neut
Marii Dąbrowskiej	Maria:Np,gen.sg.fem Dąbrowski:ADJp,gen.sg.fem
Bławatków	Bławatek:Np,gen.pl.m3
Pospolitą	Pospolity:ADJp,nom.sg.fem

Recognition of names boundaries are based on a created lexicon of grammatical base forms. In order to lemmatize them correctly we related grammatical base forms to canonical ones. A proper name is then searched for in this list of pairs, see Table 3.

In the example below we present the results of morphologic and semantic annotation for a short fragment of a dialog³. As we can see in example 3,⁴ the value

³Translation: *from Muranowska to Dragonów street*

⁴The notion *lemma* in the example (3a.) refers to grammatical base forms.

TABLE 3: A fragment of the list of PN canonical forms

grammatical base form	canonical form
Aleja Solidarność	Aleja Solidarności
Aleja Jerozolimski	Aleje Jerozolimskie
Jan Sobieski	Jana Sobieskiego
Maria Dąbrowski	Marii Dąbrowskiej
Bławatek	Bławatków
Pospolity	Pospolita
Żwirko	Żwirki i Wigury
Żwirko i Wigura	Żwirki i Wigury

of the concept id="5" is lemmatized as *Muranowska*, and is different from the grammatical base form of the word id="18" – *Muranowski*.

- (3) a. words with morphological annotation:
`<w id="17" word="z" lemma="z" POS="PreP" morph="-" />`
`<w id="18" word="Muranowskiej" lemma="Muranowski" POS="ADJP"`
`morph="gen.sg.fem" />`
`<w id="19" word="na" lemma="na" POS="PreP" morph="-" />`
`<w id="20" word="ulicę" lemma="ulica" POS="Nc"`
`morph="acc.sg.fem" />`
`<w id="21" word="Dragonów" lemma="Dragon" POS="Np"`
`morph="gen.pl.m1" />`
- b. semantic annotation:
`<concept id="5" span="word_17..word_18" attribute="SOURCE_STR"`
`value="Muranowska" />`
`<concept id="6" span="word_19..word_21" attribute="GOAL_STR"`
`value="Dragonów" />`

The adopted solution introduces a new problem. Two (theoretically even more) names can have the same grammatical base form. It is a rare situation, we have 27 ambiguous grammatical base forms in our data. For these forms, to be able to lemmatize them properly, we introduced morphological information into the lists, see example 4. For these words we take the candidate for which both grammatical base and morphological information match.

- (4) Kaliski | Kaliska | Afem
 Kaliski | Kaliskiego | Amasc

The described schema was evaluated on a set of previously unseen 26 dialogs. In this set, 281 proper names (123 different forms) were found. For this approach, lemmatization errors can only occur if someone uses a name which is not in the lexicon or when a homomorphic name form occurs. In our data, only one such case exists as *Śląskiej* is the genitive of *Śląska* or any form of an uninflected street name

(*Aleksandry*) *Śląskie*. The only error which occurred in the test set concerned an unknown name, i.e. we got *Teatr Prezentacja* instead of *Teatr Prezentacje*.

4.2 Lemmatization based on string distance metrics

Inflected forms of names differ from nominative ones, but some fragments (substrings) remain unchanged. If we have a list of all canonical name forms, the lemmatization task can be viewed as searching this list for the closest string. This solution does not require morphological analysis of particular forms, so it can be applied without enriching the lexicon with all elements of proper names.

If we formulate our task this way, the main problem is to find the best distance measure. For our experiments we used a string metrics library implemented within the work presented in Piskorski *et al.* 2007 and Piskorski and Sydow 2007. The dialogs contains 1312 proper names in different forms, but only 754 need to be lemmatized (the rest is already in base forms). Our dictionary of canonical forms contains 4293 different names.

The results of some tested metrics are given in Table 4. In this table, the first column describes how frequently we obtained exactly one correct result, the second column shows how frequently we obtained the correct result as a single answer or as the first from a list of answers. The third column tells us how many times the only given answer was correct. Next columns give the mean and the maximum of number of results. For our task, the results from the second column are important. The metrics names' suffixes gives information about metrics parameters (e.g. *sc* for *subcostfunction*) and their meaning is as follows (for details see Piskorski and Sydow 2007):

- nw: NeedlemanWunsch;
- swwa: SmithWatermanWithAffineGaps;
- ScPol: PolishSubstCost (measure taking into account Polish specific features);
- dc: dicecoefficient used;
- scswag: subcostfunction=SmithWatermanWithAffineGaps.

Some of the cited measures gave results of above 94%. One of the best measures is unsurprisingly WLCS which takes into account common substrings, but with additional weights assigned to common beginnings. Even the most standard Levenshtein measure gave quite good results.

Two types of lemmatization errors were observed. The first one occurred when two lemmas were very close to each other, eg. for WLCS measure for *Białolece* instead of *Białoleka* we got the answer *Białolecka*. Second type of errors occurred in situations in which only a fragment of a name was used, eg. *Best Malla* instead of full *Sadyba Best Malla*. In this case we sometimes got eg. *Beskidzka* as the first answer.

5 Summary

We presented the rule based proper names annotation schema for Polish transportation information dialogs. The set of hand crafted rules assigns appropriate

TABLE 4: Lemmatization based on string metrics results

metrics	ideal Acc.	corr Acc.	single Acc.	meanlen	maxlen
JaroWinkler	0.872	0.908	0.911	2.0	2
Levenshtein	0.813	0.878	0.915	2.3	6
PermutedTokens-sw-wwa-scPol-scs- swag	0.889	0.948	0.956	2.0	3
Qgrams	0.844	0.900	0.917	2.1	4
SkipGrams	0.860	0.912	0.933	2.0	3
SmithWaterman-sw-dcy-swag	0.853	0.907	0.923	2.0	3
SmithWatermanWithAffineGaps -sw-wwa-scPol-scs- swag	0.890	0.946	0.954	2.0	3
SortedTokens-nw	0.717	0.777	0.850	3.0	11
SortedTokens-sw-wwa-scPol-scs- swag	0.901	0.949	0.956	2.0	3
LongestCommonSubstrings	0.840	0.905	0.933	2.1	4
WeightedLongestCommonSubstrings	0.875	0.942	0.952	72.4	4293

concept names to the fragments of the dialog currently being processed. The results achieved make the schema practically applicable for automatic recognition of proper names contained in user queries. As the second way for solving the task string similarity metrics have been tested. This method turned out also to be quite effective giving up to 94% correct answers for some distance measures.

Both tested methods do not take into account unknown names. For our task the rule based approach was better as it simply does not change unknown names. A string distance measure approach assigns an unknown name a lemma from existing ones. To improve our results further within the rule based approach we plan to introduce some heuristics for lemmatization of unknown names.⁵

References

- J.M. ANDERSON, editor (2007), *The Grammar of Names*, Oxford University Press.
- C. BELLEI and D. MAUREL (1997), Un dictionnaire relationnel des noms propres liés à la géographie, consultés par transducteurs, *Journal des traducteurs*, 42/2:273–282.
- P. CHRISTEN (2006), A Comparison of Personal Name Matching: Techniques and Practical Issues, in *ICDM Workshops 2006. Sixth IEEE International Conference on Data Mining*.
- A. CIEŚLIKOWA, editor (2008), *Mały słownik odmiany nazw własnych*, Rytm, Warszawa.
- W. COHEN, P. RAVICUMAR, and S. FIENBERG (2003), A comparison of string metrics for matching names records, in *Proceedings of the KDD2003*.
- T. ERJAVEC and S. DŽEROSKI (2004), Machine Learning of Morphosyntactic Structure: Lemmatizing Unknown Slovene Words, *Applied Artificial Intelligence*, 18:17–41.
- J. GRZENIA (2003), *Słownik nazw własnych*, Wydawnictwo naukowe PWN.
- K. HANDKE (1998), *Słownik nazewnictwa Warszawy*, Sławistyczny Ośrodek Wydawniczy.

⁵See Piasecki and Radziszewski 2007 for description of a guesser for Polish unknown words or Erjavec and Džeroski 2004 for lemmatization of unknown Slovene words.

- A. MYKOWIECKA, K. MARASEK, M. MARCINIAK, R. GUBRYNOWICZ, and J. RABIEGA-WIŚNIEWSKA (2007), Annotation of Polish spoken dialogs in LUNA project, in *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of 3rd Language & Technology Conference. October 5-7, 2007, Poznań, Poland.*
- M. PIASECKI and A. RADZISZEWSKI (2007), Polish Morphological Guesser Based on Statistical A Tergo Index, in *Proceedings of the International Multiconference on Computer Sciences and Information Technology.*
- J. PISKORSKI and M. SYDOW (2007), Usability of String Distance Metrics for Name Matching Tasks in Polish, in *Human Language Technologies as a Challenge for Computer Science and Linguistics. Proceedings of 3rd Language & Technology Conference. October 5-7, 2007, Poznań, Poland.*
- J. PISKORSKI, M. SYDOW, and A. KUPŚĆ (2007), Lemmatization of Polish Person Names, in *ACL 2007. Proceedings of the Workshop on Balto-Slavonic Natural Language Processing 2007 Special Theme: Information Extraction and Enabling Technologies.*
- J. RABIEGA-WIŚNIEWSKA (2008), The Syntactic Structure of Polish Proper Names of Places, in *Proceedings of FDSL7*, submitted.
- J. RABIEGA-WIŚNIEWSKA and M. RUDOLF (2003), Towards a Bi-Modular Automatic Analyzer of Large Polish Corpora, in *Investigations into Formal Slavic Linguistics. Contributions of the Fourth European Conference on Formal Description of Slavic Languages FDSL IV, held at Potsdam University, November 28-30th, 2001.*

