Linguistic Suite for Polish Cadastral System

Witold Abramowicz, Agata Filipowska, Jakub Piskorski, Krzysztof Węcel, Karol Wieloch

The Poznań University of Economics
Al. Niepodległości 10, Poznań, Poland
{w.abramowicz, a.filipowska, j.piskorski, k.wecel, k.wieloch}@kie.ae.poznan.pl

Abstract

This paper reports on an endavour of creating basic linguistic resources for geo-referencing of Polish free-text documents. We have defined a fine-grained named entity hierarchy, produced an exhaustive gazetteer, and developed named-entity grammars for Polish. Additionally, an annotated corpus for the cadastral domain was prepared for evaluation purposes. Our baseline approach to geo-referencing is based on application of aforementioned resources and a lightweight coreferencing technique which utilizes string-similarity metric of Jaro-Winkler. We carried out a detailed evaluation of detecting locations, organizations and persons, which revealed that best results are obtained via application of a combined grammar for all types. The application of lightweight co-referencing for organizations and persons improves recall but deteriorates precision, and no gain is observed for locations. The paper is accompanied by a demo, a geo-referencing application capable of: (a) finding documents and text fragments based on named entities and (b) populating the spatial ontology from texts.

1. Introduction

During the introduction of the cadastral system in Poland a lot of effort will be put on populating the core system with values of real estates. For facilitating precise value estimations one may use additional information extracted from vast amount of free-text documents, which are transmited daily via online media. This involves proper name recognition (mainly geographical references), their disambiguation and further matching documents and objects in cadastral database.

Geo-tagging of free-text data has been addressed by several research groups. Utilization of gazetters is one of the most frequently used methods (Amitay, 2004; McCurley, 2001; Pouliquen et al., 2004). (McCurley, 2001) also utilizes such information like zip codes, telephone numbers and IP-addresses of servers in order to identify the place the document refers to. (Pouliquen et al., 2004) focuses on geotagging of multilingual documents, including the ones with complex declension system like e.g. Finnish, Russian, and visualization of references on maps. Hybrid techniques for disambiguation of reference names based which combine lexical grammars and graph search on minimum spanning trees were addressed in the work of (Li, 2002).

This paper elaborates on creation of basic linguistic resources for a geo-referencing component for Polish, a lesser studied language. In particular, an exhaustive gazetteer, a named-entity grammar and an annotated corpus for the cadastral domain have been produced. Although the aforementioned resources were prepared for the cadastral domain, they are generic to some extent and can be reused for solving other information extraction tasks dealing with Polish.

The rest of this paper is organized as follows. Firstly, in section 2, we shortly discuss the issues of developing information extraction tools for Polish. Section 3 presents an overview of all created language resources in the the context of Cadastral Information System (CIS) and an evaluation of some of them. Subsequently, in section 4, we present a geo-referencing application built on top of the resources described in section 3. We finish with some conclusions in section 5.

2. Information Extraction for Polish

Polish is highly inflective and irregular language. It exhibits relatively free word order which complicates the task of information extraction (IE). Even preparation of gazetteer resources involves utilization of the complex declension paradigm (7 cases) in order to produce all variants. Further, lemmatization of named-entities is not trivial since adjectives may stand before or after a noun which poses complicacies in the process of computating the inner structure of a named-entity consisting of complex noun phrases. Last but not least, corefferencing is not only the issue of resolving name aliases and pronouns and noun phrase lemmatization, but also has to deal with zero-anaphora (e.g., in Polish it is common not to include a subject in a sentence as this information can be derived from a specific verb form). Due to the aforementioned pecularities, a grammar-based IE approach for Polish relies heavily on morphological and sentence structure features.

Relatively few research on IE for Polish has been reported. In (Piskorski, 2005a) a rule-based approach for named-entity has been presented. (Mykowiecka et al., 2005) reports on a system for content extraction from mammogram reports for Polish. Further, only very few general electronic linguistic resources for Polish exists, not to mention that only a part of them is freely available for reseaerch purposes. The work presented here relies heavily on the resources described in (Piskorski, 2005a).

3. Linguistic Suite

The suite of linguistic resources for geo-referencing is built on top of SProUT (Drożdżyński et al., 2004), a novel shallow NLP platform consisting of a pool of linguistic processing resources (tokenizer, gazetteer checker, morphology,

sentence boundary detection, and partial coreference resolver) and a grammar interpreter, where the grammar formalism is a blend of efficient finite-state devices and expressive unification-based paradigm. We have adapted and extended the existing resources for Polish, including type hierarchy, gazetteer, and named-entity grammars, described in (Piskorski, 2005a) to meet the needs of IE tools for the cadastral domain. Additionally, an annotated corpus and a spatial ontology for Polish was developed.

3.1. Named-Entity Hierarchy

We have defined a fine-grained named-entity (NE) hierarchy, which is a blend of results of previous endeavours in this area, including the work on NE taxonomies presented in (Sekine et al., 2002; Chinchor, 1998; Doddington et al., 2004). The NE hierarchy was utilized to define the DECADENT (Detecting Cadastral Entities) task on automatic detection of mentions of CIS-relevant entities in source free-text data (Filipowska et al., 2006). DECADENT focuses on recognition of entities which are explicitly referenced by their names or by a subset of nominal constructions consisting of a common noun phrase followed by a proper name. While DECADENT resembles more the MUC NE task (Chinchor, 1998), the NE categories are more similar to the categories defined in the Entity Detection Task (EDT) introduced in the ACE Program (Doddington et al., 2004). The latter task is far more complex since it requires detecting entity mentions of any type and grouping them into full coreference chains, which is beyond the scope of our current work.

There are four main categories in DECADENT task: locations, organizations, persons and products. Clearly, location is the most structured category. It groups together entities, which are relevant for geo-indexing and can be mapped onto geographical coordinates (e.g. facilities, water bodies, land forms, etc.). Particularly, administrative subcategory conforms to the geo-political division of Poland, which is organized into provinces, counties, and communes. The category product is motivated by the fact that product names often include valuable clues such as brand and company names, which can be utilized for inferring organization names and implicit location names. The category person groups named mentions of persons that are identified only via their first and/or second names or people named after a country. Each main category is subdivided into eventually non-disjoint subtypes. An excerpt of the instantiated entity hierarchy is given in figure 1.

Detecting named entities in DECADENT task consists of assigning each name mention in the source document one or possibly more tags corresponding to the type of the mentioned entity, which is accompanied by positional information. For preparation of a corporus for the DECADENT task, annotation guidelines have been prepared. The three main issues they address are: entity type ambiguity, specification of name mention borders, and producing inner bracketing of the matched text fragments.

With regard to type ambiguity problem, we primarily consider the context of whole documents in order to disambiguate the type (most of the type ambiguities in the cadastral domain concern organizations and facilities). Only,

```
LOCATION
  ADMINISTRATIVE
    CITY  -  Warszawa
    COMUNE  -  gmina Warszawa Centrum
    COUNTY  -  powiat gnieźnieński
    COUNTRY  -  Polska
    PROVINCE  -  woj. wielkopolskie
    DISTRICT  -  Rataje
    ZONE  - Nowosolska Strefa Przemysłowa
  ADDRESS
    COORDINATES  -  23 S 34 W
    STREET  -  ul. Dąbrowskiego 42
    URL  -  http://www.archive.org
    ZIP  -  61-960 Poznań
  FACILITY
    TRANSPORTATION HUB  -  Poznań Główny
    TRANSPORTATION ROUTE  -  most Św. Rocha
    UTILITY  -  Stary Browar
    ENTERTAINMENT  -  pomnik Rejewskiego
  LAND FORM  -  Dolina Nidy
  WATER BODY  -  Kanał Ulgi
ORGANIZATION
  COMMERCIAL  -  Elektromontaż Poznań
  EDUCATION  -  Uniwersytet Mikołaja Kopernika
  GOVERNMENT  -  Urząd Miasta w Tczewie
  HEALTH  -  Szpital Powiatowy w Braniewie
  RECREATION  -  KKS Lech Poznań SA
  OTHER  -  Unia Europejska
PERSON  -  Witold Gombrowicz
PRODUCT  -  Gazeta Wyborcza
```

Figure 1: Named Entity Hierarchy

in case of uncertainty, multiple annotations are allowed.[1] With respect to subtypes, we decided to assign the most specific tag as far as possible, which is similar to EDT annotation guidelines (Doddington et al., 2004).

Specification of what actually constitutes a name mention in Polish may be somewhat problematic, e.g., frequently location PPs (e.g. *w Poznaniu*) may constitute a part of a full name of an organization (e.g., *Akademia Ekonomiczna w Poznaniu* - The Poznań University of Economics). In cases, where it is not clear, PPs are detached from organization names. Further complicacies are caused by common noun-phrase keywords, which might occur either written in lowercase form or with initial capitals, and are potentially a part of the name. In general, we always handle such NP keywords written in lowercase as a part of the name mention, if removing them from the text changes the entity type (e.g. *most Św. Rocha* - bridge vs. *Św. Rocha* - person or eventually a street name). Consider *rzeka* (river) in *rzeka Odra* as a counter example. Actually, most of the nominal keywords in case of location names are treated as part of the names.

Once name mention boundaries are identified, some internal bracketing reflecting the inner structure of the complex name mentions should be produced, namely mentions including mentions of other entities. Consider as an example, the inner bracketing of [*Szkoła Podstawowa im.* [*Kornela Makuszyńskiego*] *nr. 80 w* [*Poznaniu*]] (Primary School, named after Kornel Makuszyński, in Poznań). Intuitively, only annotations of 'inner' entities which are related to CIS and geo-referencing would be produced, but for the sake of completeness, integrity and potential utilization of the

[1] We strive to solve as many type ambiguities as possible while annotating the corpus, since unambiguous information is highly relevant for automatic learning of animacy of named entities, which is a feature heavily utilized in coreference resolution approaches.

annotated data (e.g., automatic induction of NE-grammar rules, evaluation of recognition of entities of a single type, and learning type disambiguating clues), all inner entities (with some exceptiopns) are annoted. Please refer to (Filipowska et al., 2006) for more details.

3.2. Annotated Corpus

For evaluation purposes and proximate research area on automatic learning of patterns for NE recognition and relation extraction, we have created an annotated corpus which conforms to DECADENT guidelines (Filipowska et al., 2006). It consists of articles from three different sources: (a) the real estate supplement to the on-line version of Polish daily newspaper *Rzeczpospolita* (RZ), (b) the online financial magazine *Tygodnik Finansowy* (TF), and (c) different local news portals (NP) which provide news concerning events centered around development of urban architecture in Poland.

The corpus contains circa 5% of overlapping annotations, i.e. nested mentions. An overview of corpus statistics is given in table 1.

Source	Volume (kB)	#doc	#words	#tags	#tags/#doc
RZ	193	25	26750	1506	60,24
TF	180	100	23247	1689	16,89
NP	80	31	10765	875	28,23
Total	453	156	60762	4070	26,09

Table 1: Corpus statistics

3.3. Gazetteer

Gazetteers are heavily used in the context of IE. Therefore, a significant number of new entries related to locations were added to the existing gazetteer. For part of the entries, morphological variants have been produced and tailored to the modified hierarchy of named-entity hierarchy. This was done in a semi-automatic manner by application of general inflection patterns and manual correction of erroneous entries. Since extensive gazetteers are indispensable, a new compression technique for storing gazetteers has been developed, which is described in detail in (Piskorski, 2005b). Table 2 summarizes the gazetteer resources. The type of entries for which morphological variants were created are marked with an asterisk. While most of the entries are for Polish, some English names have been incorporated as well, since they appear in Polish texts frequently. Further, location entries are typically enriched with additional attribute that reflects the administrative division of the country.

According to (Li, 2002), in Tipster Gazetteer a location entry has on an average 1,39 senses and around 19% of the location entries have at least one meaning. In Polish over 12,9% of city name variants are morphologically and semantically ambiguous. For instance, there are 70 villages named *Zalesie*. Another complicacy is caused by the name convention for counties. 21,1% of county names are also city names (urban counties), where each of those cities is also a capital of land county whose name is an adjectival form of a city. The most common ambiguity types in our gazetteer are summarized in the table 3.

Type	PL	ENG
city*	155810	1006
commune	2489	-
county*	375	-
province	16	-
country	1763	-
region*	488	52
river	283	43
sea	69	-
lake	48	-
zone	23	-
facility*	68	-
given-name*	1796	16714
surname	-	13376
position*	530	-
facility	68	-
comapny	262	91
org-government	60	83
org-education	21	1276
org-recreation	12	-
org-other	119	-
other	402	-
total	164634	19265

Table 2: Gazetteer Resources

Ambiguity type	frequency
city-commune	2082
city-given-name	204
county-city	61
county-commune	56
county-city-commune	55
city-river	44
city-country	33
city-region	21
comapny-city	19
comapny-city-commune	13

Table 3: Gazetteer Type Ambiguities

3.4. NE Grammars

Obviously, applying solely gazetteer does not guarantee high coverage and might result in spurious ambiguities. Therefore, we developed NE grammars which take a broader context into account. We have adapted the general grammars described in (Piskorski, 2005a) for our task. Consider for instance street names. Storing all street names in the gazetteer would be space inefficient, laboriuos and inflexible (changes in names, typos), therefore dedicated NE grammars were created. Addresses in Polish may be relatively easily distinguished: they start with designation (one of *ul., al., pl., Oś.* or its full form), followed by a core name (e.g. person in genitive, organization name in genitive, adjective or date) and optionally followed by a numeric or alphanumeric string. A sample SProUT rule for detecting street names is gigen below. See (Drożdżyński et al., 2004) for the detailed specification of the formalism.

```
pl_ulica_bez_numeru :>
   morph & [STEM "ulica", CSTART #cs] |
   ( ( token & [SURFACE "Ul", CSTART #cs] |
      token & [SURFACE "ul", CSTART #cs] )
    token & [TYPE dot] )
   @seek(pl_geo_all_names) & [SURFACE #name, CEND #ce]
->  ne-location-postal &
     [LOCTYPE fac, LOCSUBTYPE trr,STREET #name,
  LOCNAME #all,NCSTART #cs, NCEND #ce],
where #all=ConcWithBlanks("ul.", #name).
```

SProUT grammars can be recursively embedded. We utilize this feature, e.g., person and date grammars are used for location extraction. Further, we use use only the "longest

match" strategy which enables us to disambiguate NE categories, i.e. locations from persons (e.g., *ul. Adama Mickiewicza* - Adama Mickiewicza Street), organizations from persons (e.g., *Uniwersytet Mikołaja Kopernika* - Mikołaj Kopernik University), and locations from organizations (e.g. *Rondo ONZ* - ONZ Roundabout).

Currently, the extended NE grammars for Polish cover: organizations (33 rules), persons (15), geographical and geopolitical names (51) and other entities not considered in DECADENT task like: numerical expressions (18), measurements (26), time expressions (50), other auxiliary rules (5). However some part of the latter are used as subgrammars in the geo-referencing task. The grammars are constantly extended.

Detection of some multiword names solely via utilization of NE-grammars and gazetteers may be hard due to missing contextual clues. In order to increase the coverage we procede as follows. Firstly, we select in text all sequences of uppercase tokens that were not consumed by the NE-grammar interpreter (candidates for NEs). Subsequently, for all such candidates, we compute a *Jaro-Winkler* string-based similarity, a surprisingly good and fast edit-distance metric (Cohen et al., 2003), to other previously identified names. *Jaro-Winkler* metric is an extension of the *Jaro* distance metric, which is computed as follows. Let for two strings s and t, s' be the characters in s that are common with t, and let t' be the charcaters in t that are common with s (a character a in s is *in common* with t if the same character a appears in about the place in t). Let $T_{s,t}$ measure the number of transpositions of characters in s' relative to t'. The *Jaro* similarity metric for s and t is given by equation 1.

$$J(s,t) = \frac{1}{3} \cdot \left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{2|s'|} \right) \quad (1)$$

The *Jaro-Winkler* extension modifies the weights of poorly matching pairs s, t that share a common prefix. The output score is given in equation 2.

$$JW(s,t) = J(s,t) + (PL * PS * (1.0 - J(s,t))) \quad (2)$$

In the above equation, PL dentotes the legth of common prefix at the start of the string, and PS is a constant scaling factor for how much the score is adjusted upwards for having common prefix's. In our experiments $PS = 1.0/(80\% \cdot max(|s|,|t|))$ Pairs of names whose distance are less than a given predefined treshold are grouped into correfering expressions.

Finally, we also apply additionally acronym building patterns. In this manner additional mentions of entities recognized previously by the grammars are detected.

In order to evaluate our baseline approach to NE detection, we have carried out several experiments on the corpus described in section 3.2. Firstly, we have evaluated detection of each entity type separately, by applying only specific grammar for this task. Secondly, we have merged the single grammars into one and performed detection of all entities simultaneously. Further, we have enhanced the aforesaid experiments with an application of the lightweight coreference resolution technique mentioned earlier in this section.

The evaluation results for detecting locations, organizations and persons are given in table 4, 5, and 6 respectively. The first column specifies the configuration of the experiment, where: L, O and P means application of location, organization and person grammar respectively, and C means application of lightweight coreferencing. In the other columns, *c* stands for the number of correctly identified entities (type is identified correctly and the entity borders are exactly matched), *n* means number of entities in the corpus which were not detected, *i* stands for the number of identified entities which are missing in corpus (false positives), and finally *p* stands for the number of correctly identified entities, where the entity borders are matched only partially. Consequently, the columns labeled with PP, RP, PE, and RE give the numbers on precision for partial match, recall for partial match, precision for exact match, and recall for exact match respectivelly. The aforementioned metrics has been computed as follows: $PP = (c+p)/(c+p+i)$, $RP = (c+p)/(c+p+n)$, $PE = c/(c+p+i)$, and $RE = c/(c+p+n)$.

Config	c	n	i	p	PP	RP	PE	RE
L	1111	517	428	151	0.75	0.71	0.66	0.62
LC	1126	490	553	163	0.70	**0.72**	0.61	**0.63**
LOP	1065	565	264	149	**0.82**	0.68	**0.72**	0.60
LOPC	1076	545	301	158	0.80	0.69	0.70	0.60

Table 4: Evaluation - locations

Config	c	n	i	p	PP	RP	PE	RE
O	367	843	65	240	0.90	0.42	**0.55**	0.25
OC	469	695	157	286	0.83	**0.52**	0.51	**0.32**
LOP	356	858	50	236	**0.92**	0.41	**0.55**	0.25
LOPC	433	750	91	267	0.88	0.48	**0.55**	0.30

Table 5: Evaluation - organizations

Config	c	n	i	p	PP	RP	PE	RE
P	334	124	53	36	0.87	0.75	0.79	0.68
PC	339	110	72	45	0.84	**0.78**	0.74	**0.69**
LOP	322	137	36	35	**0.91**	0.72	**0.82**	0.65
LOPC	327	125	40	42	0.90	0.75	0.80	0.66

Table 6: Evaluation - persons

As can be observed, the application of a combined grammar (LOP) improves precision in all three cases, whereas recall decreases slightly. The application of lightweight coreferencing (C) for organizations and persons improves recall, but deteriorates precision. In case of locations no gain could be observed. Probably more data is needed in order to estimate thoroughly the impact of lightweight coreferencing. A detailed analysis yielded that most errors were mainly caused by: (a) omission of some crucial keywords in the grammars, in particular in the case of organizations, (b) ambiguities and missing morphological variants in the gazetteer (some part of the grammar rules rely only on gazetteer data), (c) annotation errors (e.g., clashes between locations vs. organizations), (d) type ambiguities due to the lack of context knowledge (e.g., *powiedział prezes spółki*

Kruk - [said CEO company$_{gen}$ Kruk], where *Kruk* might either refer to a person (CEO) or a company, even if we consider the subcategorization frame for the verb *powiedzieć* (to say).

A detailed evaluation for subtypes for the LOP and LOPC configuration is given in table 7. The numbers in brackets correspond to the the results with LOPC configuration. As expected the application of the *Jaro-Winkler* string-similarity metric for coreferencing improves the recall, especially in the case of organizations. However, the numbers in the table indicate that more data is needed in order to estimate its usefullness.

Subtype	PP	RP	PE	RE
loc-adm-cit	0,56 (0,55)	0,75 (**0,76**)	0,55 (0,53)	0,73 (0,73)
loc-adm-cmn	1,00 (1,00)	0,40 (0,40)	1,00 (1,00)	0,40 (0,40)
loc-adm-cry	0,94 (0,91)	0,93 (0,93)	0,94 (0,91)	0,93 (0,93)
loc-adm-pro	0,96 (0,93)	0,90 (0,90)	0,96 (0,93)	0,90 (0,90)
loc-adr-str	0,90 (0,87)	0,73 (0,73)	0,87 (0,84)	0,70 (0,70)
loc-adr-url	1,00 (1,00)	0,08 (0,08)	0,50 (0,50)	0,04 (0,04)
loc-fac-ent	0,56 (0,50)	0,39 (0,39)	0,41 (0,36)	0,29 (0,29)
loc-fac-trh	0,71 (0,71)	0,46 (0,46)	0,41 (0,41)	0,27 (0,27)
loc-fac-trr	0,76 (0,75)	0,45 (0,45)	0,71 (0,70)	0,42 (0,42)
loc-fac-uti	0,44 (0,39)	0,15 (0,15)	0,35 (0,32)	0,12 (0,12)
loc-lan	0,84 (0,76)	0,64 (**0,67**)	0,84 (0,73)	0,64 (0,64)
loc-wat	0,43 (0,43)	0,25 (0,25)	0,43 (0,43)	0,25 (0,25)
org-com	0,92 (0,89)	0,35 (**0,43**)	0,43 (**0,46**)	0,16 (**0,22**)
org-edu	0,86 (0,84)	0,58 (**0,64**)	0,45 (**0,48**)	0,30 (**0,36**)
org-gov	0,92 (0,90)	0,47 (**0,48**)	0,81 (0,80)	0,41 (**0,43**)
org-hlt	0,50 (**0,80**)	0,13 (**0,50**)	0,50 (0,20)	0,13 (0,13)
org-oth	0,86 (0,81)	0,57 (**0,60**)	0,79 (0,75)	0,52 (**0,55**)
org-rec	0,19 (0,18)	0,21 (0,21)	0,17 (0,15)	0,18 (0,18)
per-nam	0,91 (0,90)	0,72 (**0,75**)	0,82 (0,80)	0,65 (**0,66**)

Table 7: Evaluation - subtypes

3.5. Ontology Population

The created gazetteer and grammar resources are not very useful outside SProUT. Many applications (e.g. searching, navigation, visualization) require ontologies to present user with extracted named entities. The ultimate goal of applying grammars is to build spatial ontology representing knowledge acquired from the analyzed documents. The annotation of documents with underlying ontology and extraction of instances is referred to as ontology population (Amardeilh et al., 2005).

Gazetteer and grammars are complimentary in a sense that gazetteer provides basic structure for spatial ontology and grammars allow to extract instances for the ontology. Grammars are useful for recognition of names with some designations, e.g. street names (prefix *ul.*), company names (suffix *S.A.*), people (prefix *Prof.*). When such designation lacks we put the name in the gazetteer instead of using some higher level heuristics. Some of the entities are recognized purely by gazetteer (e.g. cities), by grammar (e.g. streets) or both (e.g. lakes: *Wigry* vs. *Jezioro Mikołajskie*).

Results of extraction by gazetteer and by grammar should be merged. We have assigned groups of grammar rules to the appropriate types in the type hierarchy, e.g. rules recognizing street names to LOC-ADR-STR. While locations in the gazetter are hierarchicaly related, streets recognized by grammars definitely lack this information. Placing instances recognized by the grammar under appropriate instances recognized by gazetteer requires taking into account contextual information. For example, in order to assign a street to an appropriate city we use two heuristics. If there is a prepositional phrase indicating directly the city name, we use this city name, e.g. *ul. Wysoka w Krakowie*. Otherwise we use the last city mentioned in the document.

4. Geo-referencing System

A geo-referencing application has been built on top of the resources described in previous section. Spatial ontology populated during analysis of documents may be used for searching and navigating in the corpus. User may select concepts in order to find documents or text fragments. In order to faciliate the selection, concepts are displayed as a tree showing the relations between them. As there may be several relations in the ontology, it should be projected on tree structure according to one of the relations. The most important relation in spatial domainis "part of", i.e. showing location of one entity within another entity, however also other relations may be considered, e.g. "north of", "neighbour of". In the developed application we use LOC-ADM hierarchy, with country as a root and streets as leaves. A screenshot of the GUI is depicted in figure 2.

On collection level, the projected ontology allows to select documents that are annotated with a given entity (e.g. a city) or related entities (e.g. commune with its all cities). Shall a user select another concept or a concept from another domain ontology, the resulting document collection is further constrained. In this case ontologies provide help in navigation.

On document level, by selecting a concept or an instance in a tree, all of its occurrences in the viewed document are highlighted. Moreover, taking into account concepts on different levels it is possible to structure the document into separate contexts, independent from the syntactical structure (i.e. paragraphs or sentences). For example, it is important to know which fragments of a document concern a given city, because then it is possible to ascribe a mention of a street name to this city. The same applies for non-unique cities and communes or counties.

As to the application design, user should also have the possibility to correct annotation or add omitted one, however this feature has not been introduced yet.

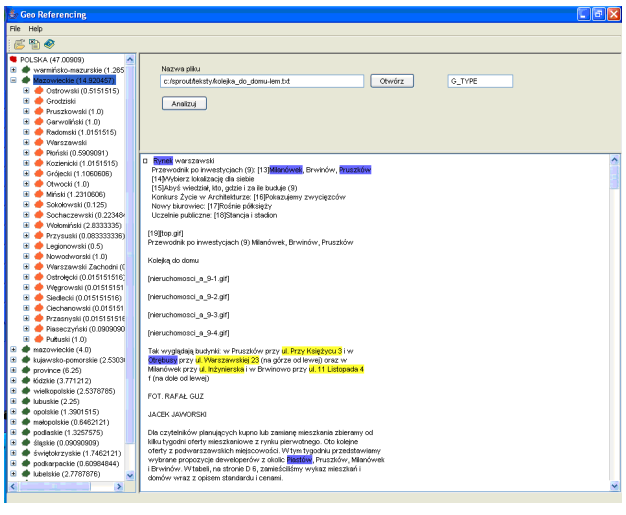

Figure 2: GUI

Matching documents and cadastral objects is out of the scope of this paper; however, this task is reduced to finding mappings between spatial ontology extracted from documents and hierarchy encoded in the cadastral system.

5. Conclusions and Future Work

In this paper we have presented an effort of creating basic linguistic resources for geo-referencing of Polish free-text documents. In particular, we have created a fine-grained named entity hierarchy, an exhaustive gazetteer, a pool of named-entity grammars for Polish, and an annotated corpus for the cadastral domain. Further, we have provided some evaluation of the coverage of these resources and we have implemented a basic geo-referencing tool on top of them. The results are fair, but there is a lot of space for improvement since we primarily deploy well-known techniques. Hence, the presented resources constitute a baseline for our proximate research. Although there are some linguistic resources for Polish, they were not suitable for utilization in development of cadastral information systems.
Based on the result of DECADENT task, defined in this paper, we will define the next-level task DEMENTI —- DEtecting Mentions of ENTIties, which will address detecting of entity mentions of any type, including nominal and pronominal mentions. Especially annotation guideline for this task will be a challenging task due to the problems described in section 2., i.e. zero-anaphora, which are very frequent in Polish.

6. Acknowledgement

This research project has been supported by a Marie Curie Transfer of Knowledge Fellowship of the European Community's Sixth Framework Programme under contract number MTKD-CT-2004-509766 (enIRaF). Part of the annotated corpus will be published on the web page of the project.

7. References

Amardeilh, F., P. Laublet, and J.-L. Minel. (2005). Document annotation and ontology population from linguistic extractions. *In Proceedings of the 3rd international conference on Knowledge capture*, ACM Press: Banff, Alberta, Canada, p. 161-168.

Amitay, E., Har'El, N., Sivan, R., Soffer, A. (2004). Web-a-where: Geotagging Web Content. *In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, Sheffield, United Kingdom, pages 273-280.

Chinchor, N. A. (1998). Overview of MUC-7. *Message Understanding Conference Proceedings.*

Cohen, W., Ravikumar, P., Fienberg, S. (2003). A Comparison of String Distance Metrics for Name-Matching Tasks. *In Proceedings of IJCAI-03 Workshop on Information Integration on the Proceedings of IJCAI-03 Workshop on Information Integration on the Web (IIWeb-03)*, Acapulco, Mexico, p. 73–78.

Day, D., McHenry, Ch., Kozierok, R., Riek, L. (2004). Callisto : A Configurable Annotation Workbench. *In Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation*, Lissbon, Portugal.

Doddington, G., Mitchell, A., Przybocki, M., Ramshaw, L., Strassel, S., Weischedel, R. (2004). Automatic Content Extraction (ACE) program - task definitions and performance measures. *In Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation*, Lissbon, Portugal.

Drożdżyński, W., Krieger, U.-U., Piskorski, J., Schäfer, U., Xu, F. (2004). Shallow Processing with Unification and Typed Feature Structures — Foundations and Applications. *In Künstliche Intelligenz*, Vol. 2004(1), pages 17–23.

Filipowska, A., Piskorski, J., Wecel, K., Wieloch, K. (2006). An Annotated Corpus for Development of Modern Cadastral Information Systems. Under submission.

Laprun, Ch., Fiscus, J., Garofolo, J., Pajot, S. (2002). A Practical Introduction to Atlas. *In Proceedings of LREC 2002: Third International Conference on Language Resources and Evaluation*, La Palma, Canary Islands, Spain.

Mykowiecka, A., Kupść, A., Marciniak, M. (2005). Rule-based Medical Content Extraction and Classification. *In Proceedings of New Trends in Intelligent Information Processing and Web Mining Conference*, Springer Verlag.

Li, H., Srihari, R.K., Niu, Ch., Li., W. (2002). Location Normalization for Information Extraction. *In Proceedings of the 19th international conference on Computational linguistics*, Taipei, Taiwan, pages 1–7.

McCurley, K. S. (2001). Geospatial Mapping and Navigation of the Web. *In WWW 10*, Hong Kong.

Piskorski, J. (2005). Named-Entity Recognition for Polish with SProUT. *Lecture Notes in Computer Science Vol 3490, 2005: Intelligent Media Technology for Communicative Intelligence: Second International Workshop - IMTCI 2004*, Warsaw, Poland.

Piskorski. J. (2005). On Compact Storage Models for Gazetteers. *In Proceedings of the 5th International Workshop on Finite-State Methods and Natural Language Processing - Lecture Notes in Artificial Intelligence*, Helsinki, Finland.

Pouliquen, B., Steinberger, R., Ignat, C., de Groeve, T. (2004). Geographical Information Recognition and Visualisation in Texts Written in Various Languages. *ACM Symposium on Applied Computing*, ACM.

S. Sekine and K. Sudo and C. Nobata. (2002). Extended Named Entity Hierarchy. *In Proceedings of the International Conference on Language Resources and Evaluation*, Canary Islands, Spain.