An Annotated Corpus for Development of Modern Cadastral Information Systems

Jakub Piskorski, Krzysztof Węcel,
Agata Filipowska, Karol Wieloch
Department of Management Information Systems
The Poznań University of Economics
{J.Piskorski; A.Filipowska; K.Wecel;
K.Wieloch}@kie.ae.poznan.pl

Abstract

Development of modern Cadastral Information Systems (CIS) requires deployment of tools for automatic estimation of real estates' value which is influenced by a number of factors. After differentiation of the factors, apropriate information on certain locations needs to be acquired. Since most up-to-date information is transmited mainly as free-text documents via online media, information extraction technology plays a key role in converting such data into valuable and structured knowledge, which faciliates automatic real-estate value estimation.

This article reports on creation of a corpus of Polish free-text documents, tagged with name mentions of CIS-relevant entities, which constitutes a core resource for development and evaluation of information extraction components used within a cadastre framework.

1. Introduction

The traditional cadastral information system (CIS) is a system containing mostly structured data about real estates (RE), e.g. location, ownership, value, etc. RE value is influenced by number of infrastructural, socio-economical and natural factors. We claim that an enormous amount of free-text documents, produced daily by diverse online media, contains valuable information and indicators on these factors, which are useful in the process of real-estate value estimation [Abramowicz et al., 2004].

The prerequisite for extracting structured knowledge from free-text document sources is automatic detection of references to objects, potentially relevant to a cadastre (e.g. locations, organizations, person names). Our work focuses mainly on development of information extraction (IE) components for processing Polish documents. Unfortunately, existing language resources for Polish are sparse and inappropriate to tackle this task (e.g. only morphologically annotated corpora are available [Przepiórkowski, 2005]). An annotated corpus with mamed entities is

indispensable to start any endavour in this area. It may be utilized in diverse ways: (a) for testing the proof-of-concept prototype of a CIS, (b) for automatic learning of patterns for recognition of entities and relations among them, which are relevant in the cadastral domain, and (c) for cadastre ontology population.

This article reports on creation of an annotated corpus for supporting development of IE tools to be utilized as submodules for automatic knowledge acquisition in CIS. In particular, we describe the DECADENT task focusing on detection of cadastre-related entities from free-text documents for Polish. Further, we discuss some corpus annotation guidelines and encountered problems. Finally, we provide some corpus statistics.

The authors are not familiar with any similar work for Polish, however we borrowed some ideas from MUC [Chinchor 1998] and ACE annotation guidelines and taxonomies [ACE], prepared for other languages and domains. Our work is also strongly related to extracting geographical references, which has been addressed in various publications [McCurley, 2001, Pouliquen et al. 2004, Amitay et al. 2004].

The rest of this paper is organized as follows. Section 2 presents the DECADENT - task centred around detecting name mentions. Subsequently, section 3 discuses the annotation guidelines and related issues. Section 4 gives an overview of corpus statistics and the annotation tool. We end up with some conclusions in Section 5.

2. DECADENT Task

DECADENT (Detecting Cadastral Entities) task focuses on detecting mentions of CIS-relevant entities in source free-text data. We consider an entity be an object or a set of objects in the real world. Entities can be referenced in a free text by: (a) their name, (b) a common noun phrase (c) a pronoun or (d) an implicit mention in elliptical constructions (e.g., in Polish, subject is often missing in clausal constructions, but it can be inferred from the suffix of a corresponding verb form). In DECADENT task, we are only interested in recognition of entities which are explicitly referenced by their names (named entities) or by a subset of nominal constructions consisting of a common noun phrase followed by a proper name. While our task resembles more the MUC NE task [Chinchor, 1998], the NE categories are more similar to the categories of the Entity Detection Task (EDT) introduced in ACE Program [Doddington et al., 2004]. However, DECADENT task is less complicated than EDT since the latter requires detecting mentions of any type and grouping them into full coreference chains, which is beyond the scope of our current work. In other words, we recognize text fragments which may refer to some objects in the real-world, but we do not tailor them to any concrete real-world objects.

Originally ACE program specified 7 basic categories: organizations, geo-political entities, locations, persons, facilities, vehicles and weapons. They were used as base for specifying the DECADENT task, i.e., we have modified and

adapted them to meet the needs of CIS applications. For instance, the categories: locations, facilities and geopolitical entities have been merged into one category – location, which represents entities that can be mapped onto geographical coordinates. Further, we added the category product, since product names often include valuable clues such as brand and company names, which can be utilized for inferring locations and might implicitly constitute a strong indicator of real estate price level etc. Currently, in DECADENT task, there are four main types of entities:

- Locations (LOC) (natural land forms, water bodies, geographical and political regions, man-made permanent structures, addresses, etc.)
- Organizations (ORG) (companies, government institutions, educational institutions, and other groups of people defined by an organizational structure)
- Persons (PER) (individuals or groups of humans)
- Products (PRD) (brand names, services, goods)

Clearly, LOC is the most structured of the entity types. Its main purpose is to group together entities, which are relevant for geo-indexing. Each main type is subdivided into eventually non-disjoint subtypes. The category LOC groups such entities like: natural land forms (LAN) (e.g. continent names, geographical regions), water bodies (WAT), facilities (FAC), addresses (zip codes, building numbers, geographical coordinates and URL's or e-mails), and administrative regions (ADM). Facilities (FAC) are further subdivided into: transportation hubs (TRH), transportation routes (TRR), entertainment facilities (ENT) and other utilities (UTI). Administrative regions are subdivided into: countries (CRY), provinces (PRO), counties (CNT), communes (CMN), cities (CIT), districts (DIS) and other zones (ZON). See table 3 for details.

Within ORG type, we distinguish commercial organizations (COM) (companies and some other private-owned institutions), governmental institutions (GOV) (related or are dealing with the administrative issues and other affairs of government and the state), schools, universities and research institutes (EDU), organizations related to health and care (HLT), institutions dealing with recreation or media (REC), and finally other (OTH) organizations that do not fit into any of the previous categories.

PER category groups named mentions of persons that are identified only via their first and/or second names. Titles, positions, etc. are not to be detected since this information is not necessarily of an interest in the context of CIS. Further, groups of people named after a country or likewise fall into this category too.

Entities of PRD type are to be detected due to their association with organizations that promotes them. We believe that such information might be useful for inference purposes at a later stage, as mentioned earlier. Currently, we only consider brand names (BRN).

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. Due to eventual type ambiguities, difficulties in specifying name

mention borders and subtleties of Polish, we have introduced some annotation guidelines described in more detail in the next section.

3. Annotation Guidelines

This section gives a short overview of annotation guidelines. In particular, there are three major issues and problems, which have to be tackled: entity type ambiguity, specifying name mention borders, and finally inner bracketing of the matched text fragments.

3.1. Type ambiguity

Type ambiguity of named-entities is a well-known problem. While, in most cases the type of the entities in our corpus happened to be unambiguous, some other pose problems. Usually ambiguities arise between: (a) organizations and persons, (b) brand names and organizations, and (c) locations and organization, where the latter type of ambiguity is crucial and most frequent in the context of CIS (see table 4). Consider as an example the following clauses:

(1) *Wojewódzki Szpital w Bydgoszczy nabył nową aparaturę ratunkową.*
(Municipal Hospital in Bydgoszcz purchased an new rescue devices)
(2) *Wojewódzki Szpital w Bydgoszczy został wyremontowany.*
(Municipal Hospital in Bydgoszcz was renovated.)
(3) *Wojewódzki Szpital w Bydgoszczy wygrał konkurs.*
(Municipal Hospital in Bydgoszcz won a competition.)

The name *Wojewódzki Szpital w Bydgoszczy* (Municipal Hospital in Bydgoszcz) in (1) refers to the authorities of the hospital (organization), whereas in (2) it refers to the building of the hospital (location). Finally, when we disregard the context of the clause (3) appears in, it is not clear whether the name refers to the building or the authorities of the hospital. We use the following rule of thumb in such a case. If the context (either preceding or succeeding sentence or paragraph etc.) allows to unequivocally interpret the type of entity hidden behind the particular name occurrence, then a single tag should be assigned. Otherwise, if the interpretation is uncertain, two or more annotations may be assigned if necessary. 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 [Evans et al., 2000].

With respect to subtypes we decided to assign the most specific tag as far as possible. Consider as an example a private commercial educational institution which falls into either ORG-COM (commercial organization) or ORG-EDU (educational institution) class. In such a case, the more specific tag has a higher

priority, i.e., ORG-EDU. This guideline is similar to the one specified in the EDT annotation guidelines of the ACE program [Doddington et al., 2004].

Ambiguities concerning tailoring particular name mentions to real-world object, e.g., there are ca. 70 cities in Poland named *Zalesie* and several companies called *POLSOFT*, are not handled within DECADENT task. Hence no attributes are produced which link text fragments to concepts.

3.2. Name Mentions Border Detection

Specification of what actually constitutes a name mention in Polish may be somewhat problematic. First of all, we apply the longest-match strategy, i.e., we take as many tokens which are potentially part of the name as possible, e.g., we treat the whole phrase *Akademia Ekonomiczna w Poznaniu* (The Poznań University of Economics) as a name mention since *w Poznaniu* is a part of the full name of the institution (this issue does not concern English). In cases, where it is not clear, we exclude such prepositional phrases including location names from being part of the organization name. Furthermore, in case of organization names, we disregard any common noun phrases written in lowercase letters, which preceded a proper name, as a part of the name (e.g., in *grupa kapitałowa Forum* - Holding Forum, only *Forum* is tagged), even if they could intuitively constitute a part of the full name. Contrary to this, in case of locations, we consider some nominal constructions, consisting of simple lowercased common noun phrases followed by a proper name as name mentions. Let us consider the the phrase *Most Św. Rocha* which is a name of a bridge. It could be alternatively mentioned in the text as *most Św. Rocha*. Without discussing the subtleties of Polish orthography w.r.t. capitalization and the style commonly used in the newspapers etc., we decided to treat both variants as name mentions as far as the leading common noun phrases is potentially a part of the full-name (as in our example). Hence, no matter if the common noun phrase keyword being a part of the name is written in lowercase letters or starts with a capital initial letter, it is always treated as a part of a name mention. Analogously, we would annotate both *zakłady Hipolita Cegielskiego* (Hipolit Cegielski plants) and *Zakłady Hipolita Cegielskiego* as a name mention.

For solving the problem of name mention borders, we use further rules:

- If a common noun or common noun phrase keyword starts with an initial capital, is not sentence initial, and is followed by a proper name, then it is always considered to be a part of the name mention (even if one would intuitively not consider it as a part of the name), e.g. the word *Grupa* in *Grupa Kapitałowa ABC* (Capital Group ABC)
- In case of addresses all keywords, e.g. *ul.*, *Al.*, *al.*, *Plac*, etc. are a part of the name mention (likewise strategy is followed for some other location subtypes)
- If deleting a lowercased common noun phrase keyword, e.g., *pomnik* in *pomnik Adama Mickiewicza* (monument of Adam Mickiewicz), results in a name (here: *Adam Mickiewicz*), which does not mach the same entity type

(which is the case in our example), then such a keyword is a part of the name mention. Constructions like: *powiat Koszaliński* (county of Koszalin), *ocean Atlantycki* (Atlantic Ocean) are further examples of this type. As a counter example, consider the keyword *rzeka* (river) in *rzeka Odra*. Here, deleting *rzeka* does not change the type of *Odra* (in the same context). Hence, the keyword *rzeka* is not treated as apart of the name mention.

3.3. Inner Bracketing

Once name mention boundaries are identified, we eventually add some internal bracketing which reflects the inner structure of the mention to some extent. Consider the following name mentions enriched with inner bracketing.

(1) [[ul. [Jana III Sobieskiego PER-NAM] LOC-FAC-TRR] 10/4 LOC-ADR-STR] (the street named after Jan III Sobieski, Polish king)
(2) [[Osiedle [Kopernika PER-NAM] LOC-ADM-DIS] 12/2 LOC-ADR-STR] (the district of buildings named after Copernicus)
(3) [Zakłady [Hipolita Cegielskiego PER-NAM] LOC-FAC-UTI] (Hipolit Cegielski plants)
(4) [Kino [Malta LOC-ADM-DIS & LOC-WAT] LOC-FAX-ENT & ORG-REC] (cinema Malta)
(5) [Giełda Papierów Wartościowych w [Warszawie LOC-ADM-CIT] ORG-COM] (Warsaw Stock Exchange)
(6) [[Kulczyk PER-NAM] Tradex ORG-COM] (company)
(7) [fabryka [Pepsi ORG-COM & PRD] LOC-FAC-UTI] (factory)
(8) [Akademia Ekonomiczna w [Poznaniu LOC-CIT] ORG-EDU] (university name)

A question arises, how to bracket a given name mention. Intuitively, one would only consider annotations of 'inner' entities which are related to CIS and geo-referencing. Hence, in our example only inner entities in (4, 6, 7) should be annotated, since they refer to existing locations relevant for geographical indexing (4, 7), a currently living person, known to be major investor in the city of Poznań (6), which is potentially relevant to CIS, or product brand name within the facility/organization name (7).

However, for the sake of completeness, integrity and potential utilization of the annotated corpus for other tasks (e.g., automatic induction of NE-grammar rules, evaluation of components for recognition of entities of a single type, and learning type disambiguating clues), all (or almost all) inner entities are annotated. The following table gives guidelines with examples for entity type combinations (outer – inner), for which inner bracketing is provided.

Table 1. Entity type combinations.

	LOC	ORG	PER
LOC	[Ul. Biała] 13 (address)	Rondo [ONZ] (United Nations roundabout)	ul. [Jana III Sobieskiego] (street)

ORG	AE w [Poznaniu] (university name)	Wydział Prawa [UAM] (Faculty of Law of UAM)	Uniwersytet [Adama Mickiewicza] (university name)
PRD	[Warka] Strong	[Microsoft] Exchange	Piwo [Heweliusz] (beer)

Some complex nominal constructions might pose difficulties while carrying out annotations. Their inner bracketing has to be done carefully. In particular, it is important to differentiate between what we consider a full name and complex noun/prepositional phrases and appositions, which might appear tricky in some context. The following two text fragments clarify the idea:

- [*Szkoła Podstawowa im.* [*Kornela Makuszyńskiego* PER-NAM] *nr. 80 w* [*Poznaniu* LOC-ADM-CIT] ORG-EDU]
- *Siedziba* [*Microsoft* ORG-COM] *w* [*Warszawie* LOC-ADM-CIT] *w* [*Polsce* LOC-ADM-CRY]

The first one happens to be a full-name of the school (with some nested names), whereas the second one constitutes a complex noun phrase consisting of one simple noun phrase followed by two simple preposition phrases, which is unlikely to be a fullname. Hence, only *Microsoft*, *Warszawie*, and *Polsce* are tagged.

4. Corpus

In order to be able to reason about real estates value, the CIS system needs to be supplied with diversity of documents from sources being monitored. Hence the annotated corpus consists of articles from 3 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. Statistics of the corpus are given in Table 2. More fine-grained data accompanied by some examples is given in Table 3.

Table 2. Corpus statistics.

Corpus	Volume (KB)	Documents	Words	Tags	Words per document	Tags per document
RZ	193	25	26750	1400	1070,00	56,00
TF	180	100	23247	1675	232,47	16,75
NP	80	31	10765	867	347,26	27,97
total:	453	156	60762	3942	389,50	25,27

Table 3. Annotation statistics and examples.

Category		Total	Examples
LOC		**1661**	
	ADM CIT	612	*Warszawa*
	ADM CMN	20	*gmina Warszawa Centrum*
	ADM CNT	1	*powiat wołomiński*
	ADM CRY	207	*Polska*
	ADM DIS	201	*Rataje*
	ADM PRO	47	*woj. wielkopolskie*
	ADM ZON	15	*Nowosolska Strefa Przemysłowa*
	ADR COR	0	*23° S 34 ° W*
	ADR STR	35	*ul. Dąbrowskiego 42*
	ADR URL	51	*www.archive.org*
	ADR ZIP	0	*61-960 Poznań*
	FAC ENT	56	*pomnik Rajewskiego*
	FAC TRH	26	*Poznań Główny*
	FAC TRR	246	*most św. Rocha*
	FAC UTI	91	*Stary Browar*
	LAN	44	*Dolina Nidy*
	WAT	9	*Kanał Ulgi*
ORG		**1441**	
	COM	1090	*Elektromontaż Poznań*
	EDU	31	*Uniwersytet Adama Mickiewicza*
	GOV	94	*Urząd Miasta*
	HLT	9	*Szpital Powiatowy w Braniewie*
	OTH	184	*Unia Europejska*
	REC	33	*KKS Lech Poznań SA*
PER		**486**	*Witold Gombrowicz*
PRD		**354**	*Gazeta Wyborcza*

The corpus annotation was carried out by four people. The documents' pool was split into two parts and assigned to a different pair of annotators. The final annotation and annotation guidelines is a result of two iterations of the process consisting of three phases: (1) definition\tuning of guidelines, (2) annotation, (3) cross-validation. It turned out that ca. 10% of all tags had to be corrected and refined after the first iteration, which reflects the complexity of the annotation tasks.

Major problems were encountered while annotating complex and nested names. Consider as an example a name of an organizational unit: [*Zakład Konserwacji Zabytków* [*Wydziału Architektury* [*Politechniki Warszawskiej* $_{\text{ORG}}$] $_{\text{ORG}}$] $_{\text{ORG}}$] ([the Unit for the Preservation of Historical Buildings and Monuments of [the Faculty of Architecture at [the Warsaw University of Technology]]]). The name of the core organization (*Politechniki Warszawskiej* – the Warsaw University of Technology) is the most relevant for CIS. As we are not interested

in recognizing names of all intermediate organizational units, we decided to create only two annotations: one for the inner-most and the other for the outer-most name. In Table 4. we give some numbers of overlapping annotations with detailed information concerning pairwise type clashes (please compare Table 1). The number of overlapping annotaions amounts to 210, which constitute 5% of total number of annotations.

Table 4. Overlapping annotations.

	LOC	ORG	PER
LOC	60	18	6
ORG	58	23	6
PRD		38	

Another issue concerned assigning two competing tags for the same text fragment (a special case of overlapping annotations). The most frequent clash ocurred between ORG-COM and FAC-UTI types. Inferring the right one regarding the context was hard (e.g. *Centrum Spotkania Kultur*).

For carrying out the annotation task we have chosen Callisto tool [Day et al., 2004] which supports linguistic annotation of textual sources for any Unicode-supported language and allows for defining user-defined domain and task specific tags. Callisto produces a standoff annotation in AIF (ATLAS Interchange Format) format. [Laprun et al., 2002]. AIF format, implemented as an XML application, offers good properties in respect with extensibility and facilitates widespread exchange and reuse of annotation data. jATLAS is a Java implementation of the ATLAS framework [jATLAS, 2003]. It's API provides methods for modifying and querying annotations as well as reading/writing them from/to AIF files. ATLAS data model employs an extremely general notion of annotation. An ATLAS annotation picks out a region of (possibly structured) text and associates structured information (represented as nested feature structures) to it. Further, AIF supports overlapping annotations which are crucial in the context of DECADENT task.

5. Conclusions

In the article we reported on an ongoing endeavour of creating an annotated corpus for supporting development of information extraction tools for utilization in a cadastre system for converting Polish free-text documents into structured data. To be more precise, we have defined a CIS-relevant entity detection task, including a fine-rained taxonomy, and we elaborated on the annotation guidelines for preparation of the corpus and discussed the subtleties of the tagging process.

At present, the annotated corpus contains 156 documents (over 60.000 words). The described work is partly supported by the European Commission under the

Marie Curie ToK "enIRaF" (IST-509766) and a sample the corpus will be available shortly at http://eniraf.kie.ae.poznan.pl.

Our proximate work will comprise of improving our current named-entity recognition machinery via utilization of the created corpus for automatic acquisition of NE patterns. Further, a higher-level information extraction tasks, i.e. coreference resolution task (DEMENTI – Detection of Mentions) are envisaged in the near future. In particular, an appropriate corpus with annotation of all types of mentions will be prepared on top of the one described in this paper. A long-term goal will focus on amalgamation of geo-referencing and time indexing techniques to track entity history.

The work is partly supported by the European Commission under the Marie Curie ToK "enIRaF" (IST-509766).

6. References

[Abramowicz et al., 2004] W. Abramowicz, A. Bassara, A. Filipowska, M. Wiśniewski. *eVREst – Supporting Estimation of Real Estate Value*. Cybernetics and Systems. An International Journal 35 (7-8), 2004 , pp. 697-708, Taylor&Francis Group.

[ACE] ACE Program - http://projects.ldc.upenn.edu/ace/ - accessed on February 10th, 2006.

[Amitay et al. 2004] Einat Amitay, Nadav Har'El, Ron Sivan, Aya Soffer Web-a-where: geotagging web content. Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, 2004.

[Callisto] Callisto - http://callisto.mitre.org/ - accessed on February 10th, 2006.

[Chinchor 1998] Nancy A. Chinchor. *Overview of MUC-7*. Message Understanding Conference Proceedings, 1998 (http://www.itl.nist.gov/iaui/894.02/related_ projects/ muc/index.html).

[Day et al., 2004] David Day, Chad McHenry, Robyn Kozierok, Laurel Riek. *Callisto : A Configurable Annotation Workbench*. In Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.

[Doddington et al., 2004]George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, Ralph Weischedel. *The Automatic Content Extraction (ACE) Program - Tasks, Data, & Evaluation*. In Proceedings of LREC 2004: Fourth International Conference on Language Resources and Evaluation, Lisbon, Portugal, 2004.

[Evans et al., 2000] R. Evans, C. Orasan. *Improving anaphora resolution by identifying animate entities in texts*. Proceedings of the Discourse Anaphora and Reference Resolution Conference (DAARC2000), Lancaster, UK, 2000.

[jATLAS, 2003]. *jATLAS, a Java implementation of the ATLAS framework.*
[http://www.nist.gov/speech/atlas/jatlas/] - accessed on February 10th, 2006.

[Laprun et al., 2002] Christophe Laprun, Jonathan Fiscus, Joh Garofolo, Sylvian Pajot. *A Practical Introduction to Atlas*. In Proceedings of LREC 2002: Third International Conference on Language Resources and Evaluation, La Palma, Canary Islands, Spain, 2002.

[McCurley, 2001] Kevin S. McCurley. Geospatial Mapping and Navigation of the Web. WWW10, Hong Kong.

[Pouliquen et al. 2004] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, Tom De Groeve. Geographical Information Recognition and Visualisation in Texts Written in Various Languages. ACM Symposium on Applied Computing, ACM 2004.

[Przepiórkowski, 2005] Adam Przepiórkowski. *The IPI PAN Corpus in Numbers*. Proceedings of the 2nd Language & Technology Conference, Poznań, Poland 2005.