

EXPLOITING TEXTUAL MENTIONS FOR THE EXTRACTION OF PERSON ATTRIBUTES

Cristina Guerrero Flores
University of Trento - UniTN
c.guerrero@studenti.unitn.it

Silvana Bernaola
Research Institute Tutor – FBK-irst,
bernaola@fbk.eu

Alberto Lavelli
University Tutor - FBK-irst,
lavelli@fbk.eu

ABSTRACT

The problem of automatically identifying entities, relations, or events within text, in order to store this information in a structured form, has been addressed by different Information Extraction techniques. Finding the names of people and places in a document constitutes an example of these tasks. For the identification of person attributes we propose an approach based on the orthographic and morpho-syntactic information obtained only from the mentions of the entities. The results presented in this report correspond to the classifiers built for the recognition of the last names, first names, professional occupations and proveniences. This work has been carried out on a collection of Italian news articles, which had been previously annotated.

Index Terms— Information extraction, attribute extraction, Entity Mention Recognition, Entity Attributes, Ontology Population

1. INTRODUCTION

One of the goals of Information Extraction (IE) is finding relevant entities, logical pieces of information, and their relationships within textual documents. An entity is an abstraction of a specific individual or object in the world, which is mentioned within a content. A mention is the way in which a textual reference to an entity appears in the text.

The term Named Entity, widely used in Natural Language Processing (NLP) applications, was introduced within the Message Understanding Conferences (MUC)¹ as it was noticed the importance of recognizing information units like person, organization and location names, and numeric and percent expressions.

Named Entity Recognition (NER)[1] is a subtask of Information Extraction whose objectives are to locate and

classify words in text into categories, such as proper names, organizations, locations, expressions of time, among others.

Another activity that takes part in the processing of contents is the Entity Mention Detection (EMD) task², which consists on the detection of the extension of entities and their classification in selected semantic classes, i.e. Person (PER), Organization (ORG), Geo-Political entity (GEO) or Location (LOC).

There are different levels of mentions² according to the way the entity reference is made; these levels are: name, nominal or pronominal. For example, given this passage of a news content “Venezuelan President Hugo Chavez called for Internet regulations. He accused a news Web site of spreading false information; all “Venezuelan President”, “Hugo Chavez”, and “He”, refer to the same entity in different levels.

Some mentions may include different elements or pieces of information. From the mention “Hugo Chavez”, we can extract Hugo as **FIRST NAME**, and Chavez as **LAST NAME**. These pieces of information are an example of person attributes.

The task of extracting attributes³ has gained popularity in the last years. Improving the accuracy on the identification of entities and related attributes on web search results is one application of this endeavor [2]. Ontology population is also a goal application of this task [3,4].

One of the solutions given to this problem has been the use of context words extracted from the entire document combined with manually provided regular expressions [5].

2 ACE English Entity Guidelines v6.6 4 2008.06.13
[http://projects ldc.upenn.edu/ace/docs/English-Entities-Guidelines_v6.6.pdf]

3 WePS Web People Search Guidelines for the WePS-3 Attribute Extraction Subtask [<http://nlp.uned.es/weps/weps-3/>]

1 MUC-7 Conference Proceedings [<http://muc.www.saic.com/>]

Our proposal is based on the possibility of exploiting Machine-Learning techniques to develop a recognizer of attributes directly from the diverse mentions of the entities.

This idea came up from two facts. The first is that every type of entity has particular characteristics, which we have called attributes. For example, table 1 lists the attributes for a PERSON entity.

The second fact taken into account is that mentions identify the different ways of describing an entity. Therefore it is logical to consider that the employment of the information within the mentions may lead to the elements of the entities from they were extracted.

Table 1: Attributes for PERSON entity

Attributes	Possible values
FIRST NAME	Ralph, Greg
MIDDLE NAME	J., W.
LAST NAME	McCarthy, Newton
NICKNAME	Spider, Pacifier
TITLE	Prof., Mr.
SEX	actress (F), actor (M)
ACTIVITY	journalist, doctor
AFFILIATION	The New York Times
ROLE	director, president
PROVENIENCE	South American
FAMILY RELATION	father, cousin
AGE CATEGORY	boy, woman
MISCELLANEA	The men with red shoes

In this report, we focus particularly on entities of type PERSON, and the attributes LAST NAME, FIRST NAME, ACTIVITY (as a combination of Activity, Affiliation and Role) and PROVENIENCE.

The extractors of attributes are built on top of the Italian Content Annotated Bank (ICAB) document collection.

The results of this work will be included in the LiveMemories Project⁴; which already takes advantage of different high-performance content processing tools developed for the Italian language, some of them included on web services like TextPro. The challenge behind LiveMemories project is to develop automatic methods for

interpreting contents and rebuilding these huge amounts of data into “active memories”.

2. DESCRIPTION OF THE TASK

The purpose of this project was to extract entities' attributes from contents in Italian language. The solution we proposed was to develop classifiers to identify which of the tokens included in a mention can be considered as attributes of persons.

The complexity of this task derives from different factors. One of them is the fact that a mention excludes part of the context of a token which may be useful for finding patterns for the identification of these attributes.

An additional factor to consider is the insufficient number of examples of uncommon events for training a model. One example of this case is the order of the LAST NAME and FIRST NAME. Normally, a last name will appear after the first name, however the opposite can also occur. Training a model to recognize exceptional cases is not an easy task.

We propose the development of models for identifying these attributes, using orthographic and morpho-syntactic information of the tokens in the mentions.

In the following section, first we describe the collection used for the training and testing process of the classifier. Then, we provide some details on the experiments developed for the implementation of the attributes classifiers.

2.1. Dataset

The ICAB [6] document collection, whose entities, mentions and attributes were manually annotated, was extracted from the local newspaper “L'Adige” from Trento. The selected news includes 525 news documents stories which belong to four different days (September, 7th and 8th 2004, and October, 7th and 8th 2004) and are grouped into five categories: News Stories, Cultural News, Economic News, Sports News and Local News.

Each mention in the collection was annotated for the person attributes: FIRST NAME, MIDDLE NAME, LAST NAME, NICKNAME, TITLE, SEX, ACTIVITY, AFFILIATION, ROLE, PROVENIENCE, FAMILY RELATION, AGE CATEGORY, MISCELLANEA and HONORARY.

A total of 11941 mentions were selected for this work. These mentions correspond to those which refer to one or

⁴ LiveMemories Project [<http://www.livememories.org>]

multiple persons; and since the purpose of the work was to identify person attributes, we have taken into account mentions which refer to proper names and nominal names.

We measured the number of mentions which included the attributes, and also the variability of their values, table 2.

Table 2: Variability of values for person attribute

Attribute	Occurrence of attribute in mentions	Different values for attribute	Variability of values
LAST NAME	5009	1904	38.01%
SEX	3117	3	0.09%
FIRST NAME	3016	668	22.15%
ACTIVITY	916	342	37.42%
PROVENIENCE	733	286	39.02%
AGE CATEGORY	290	100	34.48%
MISCELLANEA	257	251	97.67%
FAMILY RELATION	132	44	33.33%
MIDDLE NAME	109	66	60.55%
TITLE	73	21	28.76%
NICKNAME	72	41	56.94%
HONORARY	61	55	90.16%

From this initial judgment, we discarded attributes based on the number of occurrences within the corpora, and the variability of values they could take. This decision was taken mainly because these factors affect the training process of the classifiers.

For the attribute `SEX` the task consists only in categorizing the mentions as “male” or “female”, instead of extracting a value for the attribute from the mention; this constitutes a classification problem and not an extraction problem.

The attributes selected for this work were: `FIRST NAME`, `LAST NAME`, `ACTIVITY` (combination of Activity, Affiliation and Role) and `PROVENIENCE`.

For every experiment performed, the selected mentions were divided 70% and 30% for training and test sets respectively.

2.2. Architecture of the System

As input data for the system an initial selection of mentions was made based on their types, to include only single and multiple person type mentions.

Orthographic and morpho-syntactic features were extracted for the different experiments:

- Part of Speech (POS)
- Indication of Beginning or End of mention
- Capitalization
- Lemma
- Stem
- Multicharacter (1-4) prefixes
- Multicharacter (1-4) suffixes
- Regular and lemmatized gazetteers

In order to accomplish the objectives proposed, the task was divided into two general phases.

The first phase comprised building and evaluating the performance of models under controlled conditions (subsets of the dataset and a combination of features and context windows). Several models were developed in order to test different approaches:

- uni-class classifiers for the selected attributes,
- cascade classifiers, and
- multi-class classifiers.

And the second phase incorporated the use of the complete dataset established initially, using the best classifiers obtained on the previous phase. Figure 1 describes the general architecture of the solution.

2.3. Procedure

The ICAB data was replicated locally and the retrieval of mentions was performed using SQL scripts.

After tokenizing the selected mentions, the IOB tagging format was applied for the different attributes annotations.

The most frequently applied techniques for information extraction tasks are based on machine learning as for example Support Vector Machines (SVMs). SVMs were introduced in text categorization and later used for several other NLP tasks, because of its scalability to high feature dimension.

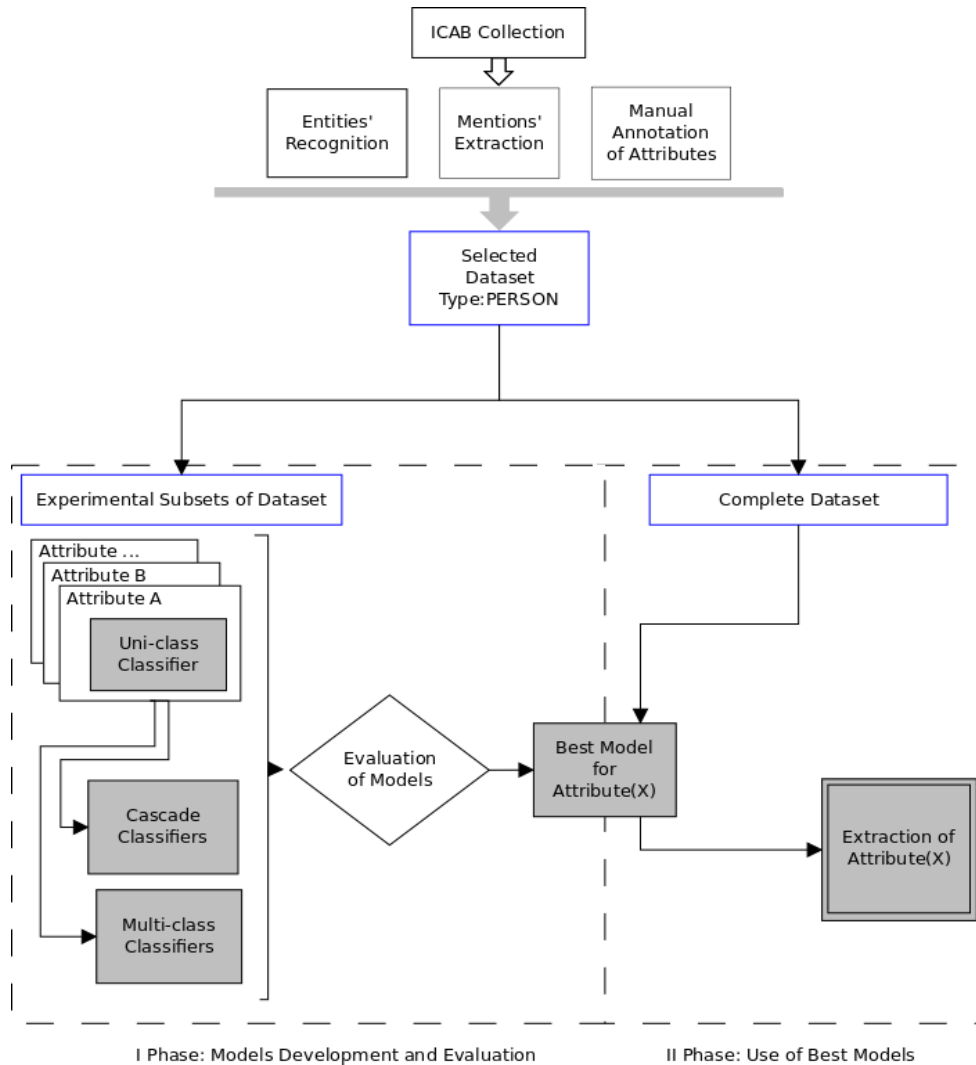


Figure 1: System Architecture

Yamcha [7] is a popular implementation and development environment of SVMs. This tool is an open source text chunker which can be customizable to other NLP tasks. YamCha allows for handling both static and dynamic features, and for defining a number of parameters such as context (window-size), parsing-direction (forward/backward) and algorithm of multi-class problems (pair wise/one vs. rest).

The attributes classifiers were built using Yamcha, a general purpose SVM-based chunker.

The features were extracted using PERL modules, and the tools TextPro [8] and Treetagger [9].

A set of initial experiments, using individually each extracted feature, revealed a number of context windows which seemed to lead us for the best models.

These windows were used on the different experiments carried out later with different combinations of features.

2.3.1 Phase I: Models Development and Evaluation

We started this phase by developing uni-class classifiers for the attributes LAST NAME, FIRST NAME and MIDDLE NAME. Although MIDDLE NAME was not among the previously selected attributes, we included it for initial experiments, as an attempt to explore the

performance of a classifier for an attribute which does not appear frequently on the data.

For this phase, three datasets were created:

- Dataset A) only positive data (only mentions which included the attribute to classify),
- Dataset B) mixed data – positive and negative samples- (all mentions from the database), and
- Dataset C) only mentions referencing a single entity of type PERSON excluding those of nominal or pronominal level.

The datasets indicated before were developed to allow us to evaluate the models on controlled and not controlled datasets. However, our idea of the best classifier to build was one which could perform well on a not controlled dataset, this means, a set of all type of mentions.

Cascade and multi-class classifiers were also developed.

2.3.2 Phase II: Application of Best Models

From the results obtained on Phase I, the best models were tested on the entire dataset described on table 2.

These results were planned as an attempt to simulate the performance of the models on a real data situation.

3. RESULTS

During the development of the task on Phase I an issue related to the manual annotation was found. Some of the initial experiments revealed mistakes on the original annotations, which lead to inaccurate results. Because of this phenomena, a new group of annotators helped on the extraction of the dataset used on the final experiments.

The baseline measures for every attribute were extracted using only gazetteers. For the multi-class classifier of the attributes LAST NAME and FIRST NAME the baseline was extracted statistically from an Italian White Pages collection. The token was tagged as LAST NAME or FIRST NAME according to the higher number of occurrences it had as a last name or as a first name within the collection.

A summary of the results achieved during the experiments performed on Phase I is shown in table 3.

After performing uni-class and cascade classifiers for the attributes LAST NAME (LNA), FIRST NAME (NAM), ACTIVITY (ACT) and PROVENIENCE (PROV), and manually analyzing the errors and results achieved from the models, we realized there was a compatibility among

the features and windows used for the extraction of attributes LNA and NAM. This characteristic of these attributes allowed us to develop a multi-class classifier, which considered LNA/NAM as a single tag. The mentioned multi-class classifier performed better than cascade and uni-class comparable classifiers.

Multi-class approaches for ACT and PROV with the attributes LNA and NAM provided results lower than the baseline; what suggested a deficient or non-existing relation among the attributes indicated.

Table 3: Results of Phase I - Experiments

Attribute	Model	Baseline	Positive samples	Mixed data
			Dataset A	Dataset B
LNA	Uniclass	91.35	95.84	45.28
	Cascade		96.43	20.62
NAM	Uniclass	90.26	97.41	43.12
	Cascade		97.51	27.84

Attribute	Model	Baseline	Mixed data	Single Person
			Dataset B	Dataset C
LNA/NAM	Multiclass	76.88	71.2	95.05

Attribute	Model	Baseline	Mixed data
			Dataset B
ACT	Uniclass	41.56	60
PROV	Uniclass	32.17	50.25

After the initial experiments performed on Dataset A we agreed that the scenario provided by that dataset was not reflecting the real conditions of a text. Therefore, results of experiments under these circumstances were not generalizable. This dataset was not used for the subsequent experiments.

In the case of PROVENIENCE, the incorporation of stem among the features increased the performance of the classifier.

The experiments with MIDDLE NAME, as expected, provided inconsistent results. The samples for training and testing were scarce, providing a poor model for recognition and misleading test results. A minimal variation in the resulting tagging lead to significant positive and negative changes on F1, but not considerable changes according to the manual evaluation.

As we explained before, once the best models were achieved from the experiments performed during Phase I, the complete original input dataset was used for a re-validation of the final classifiers. Table 4 shows the results of the performance of the best models on the final dataset, for the classifiers of the selected attributes.

Table 4: Results of Phase II - Final Classifiers

	LNA/NAM	ACT	PROV
Precision	94.06	44.8	84.52
Recall	94.45	36.66	69.18
F1	94.26	42.32	76.08

LNA/NAM: LAST NAME/FIRST NAME multi-class classifier

ACT: ACTIVITY uni-class classifier

PROV: PROVENIENCE uni-class classifier

4. CONCLUSIONS AND FUTURE WORK

From the results achieved with the experiments, we have shown that a multi-features classifier model can be effective in the extraction of attributes directly from textual mentions. The final classifiers showed an improvement over the individual attributes' baseline measures. Still some work can be made on the incorporation of non-local information in the information extraction system, or the methods for a correct combination of these attributes classifiers.

Our models take advantage of relation among attributes for the development of multi-class classifiers. An analysis of these relations among attributes of different nature could lead to the identification of new combined models.

Sets of features have been identified for the different attributes extracted; these can now be taken into account for the improvement or designing of automatic systems.

We must remark that a multi-class model for LNA/NAM, tested only on mentions referring to a single PERSON, performed 25% better than the one tested on all types of mentions as shown previously, table 3. This fact indicates that exploiting the type of mention may also help in the task of recognizing these attributes.

An unexpected result was the new set of annotations included in the source data collection. These constitute an important contribution for the improvement of a valuable resource as ICAB.

The efforts for fully realizing the task of ontology population, as presented by other authors [4], combined with the use of automatic information extraction

techniques, as the one presented in this work, constitute a step towards this undertaking. Co-reference provides another example of application which could benefit from the automatic data extracted with these models.

5. ACKNOWLEDGMENTS

This work has been partially supported by the LiveMemories project, funded by the Autonomous Province of Trento (Italy), Major Projects 2006⁴.

6. REFERENCES

- [1] N. Chinchor, E. Brown, L. Ferro, P. Robinson. (1999) Named Entity Recognition Task Definition, Technical Report Version 1.4. MITRE, Corp. and SAIC
- [2] K. Bellare, P. Talukdar, G. Kumaran, F. Pereira, M. Liberman, A. McCallum, and M. Dredze. (2007) Lightly-Supervised Attribute Extraction, NIPS 2007 Workshop on Machine Learning for Web Search, Vancouver, Canada
- [3] B. Magnini, E. Pianta, O. Popescu, M. Speranza. (2006) Ontology Population from Textual Mentions: Task Definition and Benchmark, Proceedings of the 2nd Workshop on Ontology Learning and Population (OLP2): Bridging the Gap between Text and Knowledge, Sidney, Australia
- [4] B. Magnini, E. Pianta, O. Popescu, L. Serafini, M. Speranza. (2006) From Mentions to Ontology: A Pilot Study, In Proceedings Semantic Web Applications and Perspectives (SWAP) 2006, Pisa, Italy
- [5] X. Han, J. Zhao. (2009) CASIANED: People Attribute Extraction based on Information Extraction, In Proceedings of the 2nd Web People Search Evaluation Workshop (WePS-09) at the 18th International World Wide Web Conference, Madrid, Spain
- [6] B. Magnini, E. Pianta, C. Girardi, M. Negri, L. Romano, M. Speranza, V. Bartalesi Lenzi, and R. Sprugnoli. (2006) I-CAB: the Italian Content Annotation Bank, Proceedings of Language Resources and Evaluation Conference(LREC)2006, Genova, Italy
- [7] T. Kudo, Y. Matsumoto. (2001) Chunking with Support Vector Machines, North American Chapter of the Association for Computational Linguistics (NAACL), pp. 192–199, Pittsburgh, USA
- [8] E. Pianta, C. Girardi, R. Zanolini. (2008) The TextPro tool suite. In Proceedings of LREC 2008, Marrakech, Morocco
- [9] H. Schmid. (1994) Probabilistic part-of-speech tagging using decision trees. Proceedings of the International Conference on New Methods in Language Processing (NeMLaP-1), pp. 44-49, Manchester, UK