Applying a Part-of-Speech Tagger to Postal Address Detection on the Web

Nuno Cavalheiro Marques, Sérgio Gon calves

CENTRIA   Centro de Inteligencia Artificial
Departamento de Informatica
Faculdade de Ciencias e Tecnologia
Universidade Nova de Lisboa
Portugal

nmm@di.fct.unl.pt, sergiogoncalves@netcabo.pt

Abstract

In this paper we show how a POS-tagger can be successfully adapted to a real world information retrieval system capable of extracting postal addresses from the Internet. We develop a particular tag-set for this system. Then we present and discuss the results acquired with the developed postal address tag-set. We conclude the paper by presenting a short description of the final IR system for Web postal address retrieval and drawing some conclusions.

1. Introduction

The information society is increasing its need of automatic information retrieval and extraction systems (IR). These systems should be capable of extracting knowledge from written documents, for example by using computer systems capable of word classification with tags relevant for a given task. During the last decade many part-of-speech tagging systems (POS-Taggers, e.x. (Marques and Lopes, 2001)) have been proposed. These systems are capable of ambiguity removal in syntactic tags attributed to the words in a document.

Usually part-of-speech taggers are used as preprocessors before a parser is applied. Most authors, either for comparison purposes, or because their main interest is grammatical parsing, use a tag-set derived from the original Brown corpus tag-set (e.g. the Penn-Treebank (M P. Marcus, 1993)). However for several applications this grammatical oriented tag-set is inadequate. We will present an information retrieval task, namely the extraction of postal addresses from Internet Web pages, where it was useful to develop a new tag-set.

Finally we will show how a POS-tagger that uses this tag-set can be successfully adapted to a real world information retrieval system capable of extracting postal addresses from the Internet (Goncalves, 2004).

2. A Relational Database for Postal Addresses on the Web

Following a proposal made by IBM company during last century (around the 70's), relational databases are still the most common approach to store, consult or query information on a computer. Software systems for database management are the most well known and successful area in computer science. So, today the biggest i.e. richest companies in computer Industry (such as IBM, ORACLE or Microsoft) have their own relational database systems on the market. Almost every year new and more powerful versions of the main database engines are presented.

Since the main goal behind this project was to store and query information related with postal addresses we have started by modeling the information present in a postal address. We used the traditional entity-relationship model (ER) to describe both the basic entities involved in our problem and the relations among them[1].

The mostly well-known source for postal addresses in Portugal is the Portuguese Postal Deliver Services (CTT). For research purposes, we have used some CTT data for helping us modeling an ER describing the most basic fields in a Postal Address. As it should be expected, postal addresses for a country as Portugal can be very diverse. This way, we had developed some effort in trying to model different possible Portuguese postal addresses. However we still can't pretend to have a model general to all Portuguese postal addresses. Instead we have focused on a model that could be reasonable for storing the available information in a way that could represent obvious restrictions and allowing the queries we wish to make to our system.

We have chosen to divide an address in mandatory and optional fields. As mandatory fields we have used the street identification, building specification and postal-code (or ZIP code). A postal address is only considered valid when it has all the mandatory fields. Optional fields are mainly the ones related with, e.g., the localization of an apartment like floor and apartment number.

Due to human errors, and essentially to the fact that people usually type an address over-simplifying it (normally for avoiding the use of a standardized way where we have to specify all the fields, exactly in the same annoying way), the problem of postal address detection can become quite complex for a computer (or even for a human without the right context). In fact we just can't know in advance what are the fields that are used in a given address, and most of the times two similar addresses are really the same, as in the following example:

- Rua Manuel Marques, N. 3, 6o, Dto.; 2822-343 Alenquer
- R. Manuel Marques, 3 6 dt., 2822-343 Alenquer

[1] Please see (Goncalves, 2004) for more details regarding the entity-relationship model used in this system.

Table 2. presents the basic components selected in this project for describing a postal address. This fields are distributed accross several tables in the relational system.

As it could be easily seen by reading table 2. most of these components can be directly converted into tags that can be used by a standard POS-tagger. The only exceptions are the Name of Road and Postal Region fields. The first is too complex to be directly encoded into the system. So several additional general purpose tags were considered, namely: NP - Proper Name, LIG- connector (usually an article or a preposition), NUM- A general number and $OUTRO$ - other, used for other words. Also, despite the fact that most postal regions are known, new regions can appear at any moment, so for generality we have also used the general purpose tags for this field.

Tag	Description
$TIPO$	Road Type
TIT	Title in Road Name
-N.A.-	Name of the Road
$NEDIF$	Building Number
$ANDAR$	Floor Number
FRA	Apartment Number
$POST1$	First Part of Portuguese Postal Code
$POST2$	Second Part of Portuguese Postal Code
-N.A.-	Postal Region

Table 1: Considered fields for a postal address and associated tags.

Instead of trying to develop a general semantic/syntactic tag-set, we think this tag-set should be considered as an example of a goal oriented tag-set. Indeed, at least in theory, given enough training data, all the currently available, limited context, POS-taggers could be used with this tag-set. That is why we should use this tag-set to argument in favor of more goal-oriented tagging approaches. The relation of this tag-set with more standard ones also points to the need of an algebra relating different tag-sets, namely for allowing the reuse of the tagged text in different projects.

3. The Neural Tagger

3.1. The Dictionary

In (Marques and Lopes, 2001), it was shown that a neural POS-tagger is particularly well suited for new domains where there is a lack of pre-tagged text. Since we are building a new tag-set over a new domain, this is also the case in the present problem. Also according to the results presented in (Marques and Lopes, 2001), after a minimum size for the training corpus it is the representativeness (i.e. size) of the tagger's internal dictionary and not the size of the corpus that most influences final precision of the tagging system.

Since the quality of the dictionary seems to be crucial for the tagging results, special care was taken when creating a representative dictionary for the selected tag-set. For example, numbers should be carefully treated on this problem. Indeed, the number of digits was a key property. Tag $POST1$ has 4 digits, tag $POST2$ has 3 digits. Tags $NEDIF$ and $ANDAR$ are usually smaller than 1000, so -at least in our limited training data- they also tend to have less than 4 digits. This particular domain information was taken into account by using a dictionary specific for numbers. For a given number appearing on a Web page, this dictionary converts this number into a list of probabilities that takes into account the number of digits of this number before attributing the probability for each tag (see table 3.1.).

N. Digits	Tag	Probability
1	ANDAR	53%
	NEDIF	37%
	OUTRO	11%
2	NEDIF	96%
	NUM	2%
	OUTRO	2%
3	NEDIF	14%
	NUM	05%
	POST2	81%
4	NUM	2%
	POST1	98%
5	NUM	100%
> 5	OUTRO	100%

Table 2: Values attributed by the dictionary to a number according to the number of digits.

A word dictionary was built based on the CTT data. A semi-automatic tagging procedure was applied to this dataset in order to build a good dictionary. Since the information available in CTT data is incomplete, we were forced to also include our training data to build the dictionary. This will result in a close-dictionary approach (Marques, 1999). However, according to (Marques and Lopes, 2001), these results should be similar to real ones when we increase the quality of the dictionary. Indeed table 3.1. highlights the main differences across both dictionaries.

Unknown words	A, APARTADO, C, JOAO,
	N, P, PRAC, V
Different prob.	D, DA, DE, JUNQUEIRA, TRAS

Table 3: Words appearing with different tag probabilities on both dictionaries.

In order for an idea of how hard is the proposed problem, if we classify the train set with the final dictionary we find that 19% of words have 3 possible tags, 25% words have 2 possible tags and 56% of the words have just one tag.

3.2. Adapting and Training the Neural Tagger

The neural-network tagger presented in (Marques and Lopes, 2001) was adapted for address tagging. Regarding context we have maintained the basic trigram model. So, by using dictionary lookup each sequence of three words (w_{i-1}, w_i, w_{i+1}) was converted into the three vectors with

MLE probability estimators for each possible tag (e_j) in the tag-set (E):

$$p^*(e_j|w_i), \forall j \in E.$$

These input vectors are presented to a standard feed-forward 1 layer neural-network, where each output unit was associated with one of the tags presented in the previous section. The system was trained with standard and momentum backpropagation algorithms. The data-set included only 81 postal addresses (i.e. 687 hand-tagged addresses). We have split this data-set into a training-set, validation-set (used to prevent over-fitting in the neural network) and test-set. We trained the neural-network using the ten-fold cross-validation method.

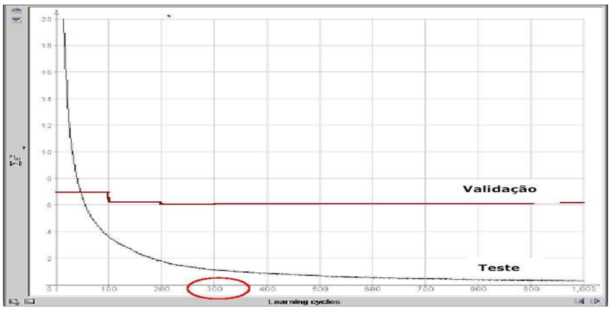
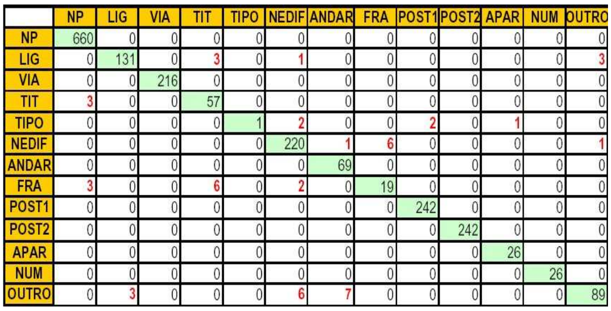


Figure 1: Evolution of the error rate during the neural-network training in the train and evaluation set. 300 iterations were enough for training this network.

Global precision rate points to a precision on the order of $97\% \pm 1, 9\%$ (for a 0.95% confidence interval). In figure 3.2. we present the cross-validation matrix resulting from the average values acquired over the ten test-set resulting from the cross-validation method.



	NP	LIG	VIA	TIT	TIPO	NEDIF	ANDAR	FRA	POST1	POST2	APAR	NUM	OUTRO
NP	660	0	0	0	0	0	0	0	0	0	0	0	0
LIG	0	131	0	3	0	1	0	0	0	0	0	0	3
VIA	0	0	216	0	0	0	0	0	0	0	0	0	0
TIT	3	0	0	57	0	0	0	0	0	0	0	0	0
TIPO	0	0	0	0	1	2	0	0	2	0	1	0	0
NEDIF	0	0	0	0	0	220	1	6	0	0	0	0	1
ANDAR	0	0	0	0	0	0	69	0	0	0	0	0	0
FRA	3	0	0	6	0	2	0	19	0	0	0	0	0
POST1	0	0	0	0	0	0	0	0	242	0	0	0	0
POST2	0	0	0	0	0	0	0	0	0	242	0	0	0
APAR	0	0	0	0	0	0	0	0	0	0	26	0	0
NUM	0	0	0	0	0	0	0	0	0	0	0	26	0
OUTRO	0	3	0	0	0	6	7	0	0	0	0	0	89

Figure 2: Confusion matrix resulting from the compilation of the several test-sets that result from the ten-fold cross-validation method. Columns correspond to expected tags and lines to the tags resulting from classification.

3.3. Discussion

The analysis of tagger results shows several interesting phenomena. We used table 3.2., complemented with tagger results analysis (when needed).

Regarding proper names (NP), they show a very high precision, without false negative cases. This could be expected since the internal dictionary was built based on a very extensive base of proper names, namely all the names of roads and places in CTT data.

The words corresponding to the connection tag LIG (mainly articles and prepositions) also present some confusion with the tags TIT and $OUTRO$. For example the D in *D' A vila* could be easily confused with the initial D in doctor (tagged with TIT). In other cases (e.g. *Vila N de Gaia*), the N is tagged with the tag $OUTRO$ (used manly for other elements in web text).

Regarding tag TIT, two types of error were detected. Compound titles such as *Capitao Mor* present some problems to the tagger due to their low frequency (the neural network, opportunistically chooses to ignore this kind of low frequency tags). These titles are usually tagged as the more frequent NP. Some other problems appear resulting of a confusion among TIT, $NEDIF$ and FRA (e.g. name of the fraction, such as *apartment 2D*). We think that some of these problems could be solved with more training data and/or by carefully balancing the training data. Also the use of longer context in the tagger could easily account for some of these cases (some preliminary experiments have been done in this direction(Correia, 2001)).

The tag $OUTRO$ has also been wrongly tagged. The main confusion occurs with tags like $NEDIF$ and $ANDAR$ in addresses like *Rua 25 de Abril*. The main problem with this address is that the tag $OUTRO$ has been used to tag numbers inside important dates that make the name of the street. Probably the best solution for this kind of cases is to retag the train corpus, changing this tags $OUTRO$ to NUM.

We should also note the high success rates on tagging tags VIA, $ANDAR$, $POST1$, $POST2$ and NUM, that although ambiguous, due to their specificity in their contexts have never been wrongly tagged.

4. The Final System

We conclude the paper by presenting a short description of the final IR system for Web postal address retrieval. This will be usefully to better present the problem and to insert the discussion of possible extensions to this work presented in the conclusions.

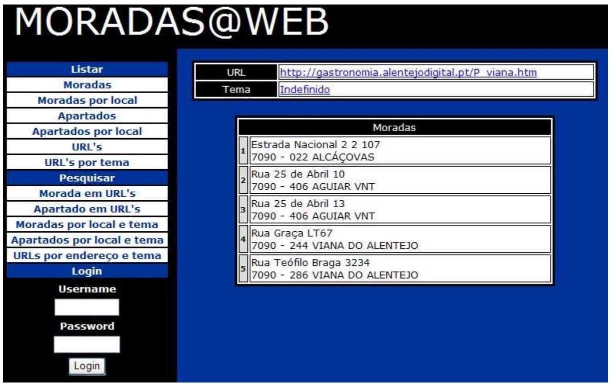
The first module of the implemented system is a Web crawler. This program is targeted to a given URL and then recursively selects the links presented on the downloaded web page to target new URLs.

Each targeted Web page is downloaded and tagged using the previously trained neural tagger. This is only possible due to the high speed of the neural tagger. Indeed, due to the nature of the neural network used, during the tagging step, all the tagging operations for a given word are reduced to a dictionary lookup and calculating the product of a vector by a matrix.

After POS-Tagging of the full text on the Web-page, a basic pattern matcher (based on regular expressions implemented in the GAWK UNIX text tool) is used to detect sequences of tags that correspond to valid postal addresses.

Whenever detected, a tentative postal address is converted into an SQL insert statement for the used database. In order to do so, we just have to use the tags previously attributed to a given address. Since the tag-set has been designed having the relational database attributes into account, the relation with the reference database becomes trivial. Also, since the tagger works on the word level, incomplete information in an address can also be easily handled. The final check on the address is performed by the integrity restrictions on the relational database (e.g. it is not possible to insert a address with an invalid postal code in the CTT data).

Figure 4. shows a screen-dump of the final system, showing several addresses located on a given web-page. Since the addresses are inserted in the database in a fully automatic way, the advantages of the relational data structure used to represent the information soon became important. For example, all the addresses related to a given site, topic or location may be listed.



URL	http://gastronomia.alentejodigital.pt/P_viana.htm
Tema	Indefinido

	Moradas
1	Estrada Nacional 2 2 107 7090 - 022 ALCÁÇOVAS
2	Rua 25 de Abril 10 7090 - 406 AGUIAR VNT
3	Rua 25 de Abril 13 7090 - 406 AGUIAR VNT
4	Rua Graça LT67 7090 - 244 VIANA DO ALENTEJO
5	Rua Teófilo Braga 3234 7090 - 286 VIANA DO ALENTEJO

Figure 3: Screen-dump showing an extracted address and the Web interface to the system.

5. Conclusions

The implemented system shows how the Neural Network POS-Tagger (Marques and Lopes, 2001), can be successfully adapted to text mining tasks in very precise domains and problems, where text analysis is needed. We think this could be an advantage to several industrial applications. We also hope to strengthen the view that tagging systems should be used not only with generic tag-sets of Part-of-speech tags, but also with tag-sets that are particular to a given task, involving -most times- very particular sub-language domains.

The integration of the text mining system with a relational database is also advantageous. The integrity constraints on the database reject several false addresses allowing for a final system with precision rates near 100%[2]). The recall rate is however low: many addresses have errors (several have been automatically detected during tests) or have incomplete mandatory fields. Further work includes the linking of the purposed system with approximate string matching techniques and data cleaning techniques. This will allow the system to automatically complete information in incomplete postal addresses.

Further applications of the purposed tagger could be made, namely on data-cleaning tasks or for feeding semantic web systems (T. Berners-Lee and Lassila, 2001).

6. Acknowledgements

The first author wishes to thank to Professor Carlos Damasio from the Computer Science Department in FCT/UNL for suggesting the treatment of this very nice problem and also for some very interesting discussions regarding a preliminary version of the work here presented.

7. References

Correia, M., 2001. Aumentando o contexto do etiquetador neuronal. Technical report, Departamento de Informatica, Faculdade de Ciencias e Tecnologia da Universidade Nova de Lisboa.

Goncalves, S., 2004. Projecto de detec cao de moradas na web. Technical report, Departamento de Informatica, Faculdade de Ciencias e Tecnologia da Universidade Nova de Lisboa.

M P. Marcus, M A Marcinkiewicz, B Santorini, 1993. Building a large annotated corpus of english: The penn treebank. *Computacional Linguistics*, 2(19):313 329.

Marques, N. C., 1999. *Uma Metodologia Estat·stica para Modela cao da Subcategoriza cao Verbal*. Ph.D. thesis, Universidade Nova de Lisboa.

Marques, N. C. and G. P. Lopes, 2001. Tagging with small training corpora. *Lecture Notes in Computer Science*, 2189:63 72.

T. Berners-Lee, J. Hendler and O. Lassila, 2001. The semantic web. *Scientific American*.

[2] This value was intuitively measured, based on our tests with the system, after the fully automatic extraction of more than one thousand postal addresses from hundreds of visited web sites.