NLP–NITMZ:Part–of–Speech Tagging on Italian Social Media Text using Hidden Markov Model

Partha Pakray
Deptt. of Computer Science & Engg.
National Institute of Technology
Mizoram, Aizawl, India
parthapakray@gmail.com

Goutam Majumder
Deptt. of Computer Science & Engg.
National Institute of Technology
Mizoram, Aizawl, India
goutam.nita@gmail.com

Abstract

English. This paper describes our approach on Part-of-Speech tagging for Italian Social Media Texts (PoSTWITA), which is one of the task of EVALITA 2016 campaign. EVALITA is a evaluation campaign, where teams are participated and submit their systems towards the developing of tools related to Natural Language Processing (NLP) and Speech for Italian language. Our team **NLP–NITMZ** participated in the PoS tagging challenge for Italian Social Media Texts. In this task, total 9 team was participated and out of 4759 tags **Team1** successfully identified 4435 tags and get the 1^{st} rank. Our team get the 8^{th} rank officially and we successfully identified 4091 tags as a accuracy of 85.96%.

Italiano. *In questo articolo descriviamo la nostra partecipazione al task di tagging for Italian Social Media Texts (PoSTWITA), che uno dei task della campagna Evalita 2016. A questo task hanno partecipato 9 team; su 4759 tag il team vincitore ha identificato correttamante 4435 PoS tag. Il nostro team si classificato all'ottavo posto con 4091 PoS tag annotati correttamente ed una percentuale di accuratezza di 85.96*

1 Introduction

EVALITA is a evaluation campaign, where researchers are contributes tools for Natural Language Processing (NLP) and Speech for Italian language. The main objective is to promote the development of language and speech technologies by shared framework, where different systems and approaches can be evaluated. EVALITA 2016, is the 5^{th} evaluation campaign, where following six tasks are organized such as:

- ArtiPhon – Articulatory Phone Recognition
- FactA – Event Factuality Annotation
- NEEL–IT – Named Entity Recognition and Linking in Italian Tweets
- PoSTWITA – POS tagging for Italian Social Media Texts
- QA4FAQ – Question Answering for Frequently Asked Questions
- SENTIPOLC – SENTIment POLarity Classification

In addition, a new challenge to this event is also organized by IBM Italy as *IBM Watson Services Challenge*. Among these challenges our team NLP–NITMZ is participated in 4^{th} task i.e. POS tagging for Italian Social Media Texts (PoSTWITA).

The main concern about PosTWITA is, Part-of-Speech (PoS) tagging for automatic evaluation of social media texts, in particular for micro–blogging texts such as tweets, which have many application such as identifying trends and upcoming events in various fields. For these applications NLP based methods need to be adapted for obtaining a reliable processing of text. In literature various attempts were already taken for developing of such specialised tools (Derczynski et al., 2013), (Neunerdt et al., 2013), (Pakray et al., 2015), (Majumder et al., 2016) for other languages, but for Italian is lack of such resources both regarding annotated corpora and specific PoS–tagging tools. For these reasons, EVALITA 2016 proposes the domain adaptation of PoS–taggers to Twitter texts.

For this task, we used a supervised leaning for PoS tagging and the details of system implementation is given in section 2. We discuss the per-

formance of the system in section 3. Finally, we conclude our task in section 4.

2 Proposed Method

For this task, we used supervised learning approach to build the model. First we implement the conditional model for PoS tagging and then to simplify the model we used Bayesian classification based generative model. Further this generative model is simplified based on two key assumptions to implement the HMM model using bigram.

2.1 Conditional Model Approach

In machine learning supervised problems are defined as a set of input called training examples $(x^{(1)}, y^{(1)}) \cdots (x^{(m)}, y^{(m)})$, where each input $x^{(i)}$ paired with a output label $y^{(i)}$. In this task, our goal is to learn a function $f : X \rightarrow Y$, where X and Y refers to the set of possible input and labels.

For PoS tagging problem, each input represents a sequence of words $x_1^{(i)}, \cdots, x_{n_i}^{(i)}$ and labels be a sequence of tags $y_1^{(i)}, \cdots, y_{n_i}^{(i)}$, where n_i refers to the length of i^{th} training example. In this machine learning each input x be a sentence of Italian language and each label be the possible PoS tag. We use conditional model to define the function $f(x)$ and we define the conditional probability as

$$p\left(y|x\right)$$

for any x, y pair. We use training examples to estimate the parameters of the model and output of the model for a given test example x is measured as

$$f(x) = arg \ \max_{y \in Y} \ p(y|x) \tag{1}$$

Thus we consider the most likely label y as the output of the trained model. If the model $p(y|x)$ is close to the true conditional distribution of a labels given inputs, so the function $f(x)$ will consider as an optimal.

2.2 Generative Model

In this model, we use the Bayes' rule to transform the Eq.1 into a set of other probabilities called *generative model*. Without estimating the conditional probability $p(y|x)$, in generative model we use the Bayesian classification

$$p(x, y)$$

over (x, y) pairs. In this case, we further break down the probability $p(x, y)$ as follows:

$$p(x, y) = p(y)p(x|y) \tag{2}$$

and then we estimate the model $p(y)$ and $p(x|y)$ separately. We consider $p(y)$ as a *prior* probability distribution over label y and $p(x|y)$ is the probability of generating the input x, given that the underlying label is y.

We use the Bayes rule to derive the conditional probability $p(y|x)$ for any (x, y) pair:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} \tag{3}$$

where

$$p(x) = \sum_{y \in Y} p(x, y) = \sum_{y \in Y} p(y)p(x|y) \tag{4}$$

We apply Bayes rule directly to a new test example x, so the output of the model $f(x)$, can be estimated as follows:

$$f(x) = arg \ \max_{y} p(y)p(x|y) \tag{5}$$

To simplify Eq.5, we use Hidden Markov Model (HMM) taggers with two simplifying assumptions. The first assumption is that the probability of word appearing depends only on its own PoS tag as follows:

$$p(w_1^n \ t_1^n) \approx \prod_{i=1}^{n} p(w_i \ t_i) \tag{6}$$

where $p(w_1^n \ t_1^n)$ means probability of tag t_i with word w_i. The second assumption is that the probability of a tag appearing is dependent only on the previous tag, rather than entire tag sequence. This is known as bigram assumption and can be measured as follows:

$$p(t_1^n) \approx \prod_{i=1}^{n} p(t_i, t_{i-1}) \tag{7}$$

Further, we incorporate these two assumptions in Eq.5 by which a bigram tagger estimates the most probable tag as follows:

$$\hat{t_1^n} = arg \max_{t_1^n} \ p(t_1^n \ w_1^n) \approx$$

$$arg \max_{t_1^n} \prod_{i=1}^{n} p(w_i \ t_i)p(t_i \ t_{i-1}) \tag{8}$$

3 Experiment Results

3.1 Dataset

For the proposed task organizers re-uses the tweets being part of the EVALITA2014 SENTIPLOC corpus. Both the development and test set first annotated manually for a global amount of $4,041$ and $1,749$ tweets and distributed as the new development set. Then a new manually annotated test set, which is composed of 600 and 700 tweets were produced using texts from the same period of time. All the annotations are carried out by three different annotators. Further a tokenised version of the texts is also distributed in order to avoid tokenisation problems among participants and the boring problem of disappeared tweets.

3.2 Results

For this task, total 13 runs were submitted 9 teams and among these runs 4 Unofficial runs also submitted. In Table 1 we list out all results for this task.

Rank	Team	Successful Tags	Accuracy
1	Team1	4435	93.19
2	Team2	4419	92.86
3	Team3	4416	92.79
4	Team4	4412	92.70
5	Team3	4400	92.46
6	Team5	4390	92.25
7	Team5	4371	91.85
8	Team6	4358	91.57
9	Team6	4356	91.53
10	Team7	4183	87.89
11	**Team8**	**4091**	**85.96**
12	Team2	3892	81.78
13	Team9	3617	76.00

Table 1: Tagging Accuracy of Participated Teams

Team 2, 3, 5 and 6 submitted one Un-Official run with compulsory one and these Un-Official submissions are ranked as 12^{th}, 3^{rd}, 7^{th} and 9^{th} respectively. We also listed these submissions in Table 1 with other runs. Our team **NLP–NITMZ** represent as **Team8** and ranked as 11^{th} in this task.

3.3 Comparison with other submissions

In this competition, a total of 4759 words were given for tagging purpose. These words were categories into 22 PoS tags and our team successfully tags 4091 words with 668 unsuccessful tags. The 1^{st} ranked team successfully tags 4435 words and the last positioned team i.e. Team9 successfully identified 3617 tags. In Table 2, we provide our system tag wise statistics.

Sl. No.	Tag	Successful Tags
1	PRON	292
2	AUX	82
3	PROPN	283
4	EMO	30
5	SYM	8
6	NUM	63
7	ADJ	145
8	SCONJ	37
9	ADP	332
10	URL	117
11	DET	288
12	HASHTAG	114
13	ADV	281
14	VERB_CLIT	10
15	PUNCT	582
16	VERB	443
17	CONJ	122
18	X	3
19	INTJ	50
20	MENTION	186
21	ADP_A	144
22	NOUN	479

Table 2: Tag wise Statistics of NLP–NITMZ Team

4 Conclusion

This PoS tagging task of EVALITA 2016 campaign is for Italian language and our system ranked 11^{th} position for the task of POS tagging for Italian Social Media Texts. We also want to mentioned that, authors are not native speaker of the Italian language. We build a supervised learning model based on the available knowledge on training dataset.

Acknowledgements

This work presented here under the research project Grant No. YSS/2015/000988 and supported by the Department of Science & Technology (DST) and Science and Engineering Research Board (SERB), Govt. of India. Authors are also acknowledges the Department of Computer Science & Engineering of National Institute of Tech-

nology Mizoram, India for proving infrastructural facilities.

References

Derczynski, Leon, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *RANLP*, pages 198–206.

Neunerdt Melanie, Bianka Trevisan, Michael Reyer, and Rudolf Mathar. 2013. Part-of-speech tagging for social media texts. In *Language Processing and Knowledge in the Web*, pages 139–150, Springer Berlin Heidelberg.

Partha Pakray, Arunagshu Pal, Goutam Majumder, and Alexander Gelbukh. 2015. Resource Building and Parts-of-Speech (POS) Tagging for the Mizo Language. In *Fourteenth Mexican International Conference on Artificial Intelligence (MICAI)*, pages 3–7. IEEE, October.

Goutam Majumder, Partha Pakray and Alexander Gelbukh. 2016. Literature Survey: Multiword Expressions (MWE) for Mizo Language. In *17th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, to be published as an issue of Lecture Notes in Computer Science, Springer. Konya, Turkey. April.