Language Segmentation of Twitter Tweets using Weakly Supervised Language Model Induction

Segmentación de Twitter Tweets con Modelos de Lenguaje Inducidos

David Alfter
University of Trier
Universitätsring 15
s2daalft@uni-trier.de

Resumen: En este artículo presentamos los primeros resultados de la inducción de modelos de lenguaje de manera semi supervisada para la segmentación por idioma de textos multilingües con especial interés en textos cortos.
Palabras clave: modelo de lenguaje, inducción, semi supervisada, segmentación, textos cortos, tuits

Abstract: This paper presents early results of a weakly supervised language model induction approach for language segmentation of multilingual texts with a special focus on short texts.
Keywords: language model, induction, weakly supervised, short text, tweet, segmentation

1 Motivation

Twitter tweets often contain non-standard language and they are limited to 140 characters. While not a problem in itself, these restrictions can pose difficulties for natural language processing systems (Lui, Lau, and Baldwin, 2014). Furthermore, Tweets may be written in more than one language (Zubiaga et al., 2014). This typically happens when multilingual speakers switch between the languages known to them, between or inside sentences (Jain and Bhat, 2014). The resulting text is said to be code-switched (Jain and Bhat, 2014; Solorio et al., 2014). This further complicates matters for natural language processing systems that need at least a certain degree of knowledge of the language at hand such as part-of-speech taggers, parsers, or machine translation (Beesley, 1988; Jain and Bhat, 2014; Zubiaga et al., 2014). The performance of "traditional" monolingual natural language processing components on mixed language data tends to be miserable, making it necessary to identify the languages in a multilingual text in order to get acceptable results (Jain and Bhat, 2014). Even if the results are not terrible, language identification and segmentation can significantly increase the accuracy of natural language processing tools (Alex, Dubey, and Keller, 2007).

Supervised methods perform well on the task of language identification in general (King and Abney, 2013; Lui, Lau, and Baldwin, 2014) and on tweets (Mendizabal, Carandell, and Horowitz, 2014; Porta, 2014), but they cannot always be applied. For one, Tweets often contain a lot of non-standard spellings and ad hoc spellings that may or may not be due to the imposed character limit. This can be problematic if the supervised methods have only seen standard spelling in training. Also, Tweets may contain languages for which there is insufficient data to train a supervised method. In these cases, unsupervised approaches might yield better results than supervised approaches.

Language segmentation consists in identifying the language borders within a multilingual text (Yamaguchi and Tanaka-Ishii, 2012). Language segmentation is not the same as language identification; the main difference is that language identification identifies the languages in a text, and language segmentation "only" separates the text into monolingual segments (Yamaguchi and Tanaka-Ishii, 2012). Language segmentation can be useful when direct language identification is not available.

2 Language Model Induction

King and Abney (2013) consider the task of language identification as a sequence labeling task, and Lui, Lau, and Baldwin (2014) a

multi-label classification task. In contrast, the proposed system uses a clustering approach. The system induces n-gram language models from the text iteratively and assigns each word of the text to one of the induced language models. One induction step consists of the following steps:

- Forward generation: Generate language models by moving forward through the text
- Backward generation: Generate language models by moving backwards through the text
- Model merging: Merge the two most similar models from the forward and backward generation based on the unigram distribution

Generation starts at the beginning of the text, takes the first word and decomposes it into uni-, bi- and trigrams. These n-grams are then added to the initial language model, which is empty at the start. For each following word, the existing language models evaluate the word in question. The highest ranking model is updated with the word. If no model scores higher than the threshold value for model creation, a new model is created.

Backwards generation works exactly the same, but starts at the end of the text and moves towards the beginning of the text.

Finally, the two models that have the most similar unigram distribution are merged. This way, the language models iteratively amass information about different languages.

The induction step is repeated at least twice. At the end of the induction, while there are two models that have a similarity greater than a certain threshold value, these models are merged.

Language segmentation is then performed by assigning each word in the text to the model that yields the highest probability for the word in question.

3 Results

Table 1 shows the results for a set of example tweets manually collected from Twitter. For all tweets, a gold standard has been manually created and evaluated against. The evaluation is that of a clustering task; the words of a text are clustered around different induced language models. Whenever the language model induction outperformed the supervised trained language models, the score is indicated in bold.

Besides the F score (F1), the F5 score is also indicated. This score sets β to 5, weighting recall higher than precision. This means that throwing together pairs that are separate in the gold standard is penalized more strongly than splitting pairs that occur together in the gold standard (Manning, Raghavan, and Schütze, 2008).

For comparison purposes, a supervised approach as described in (Dunning, 1994) has been implemented. For the supervised approach, language models for all relevant languages have been trained on Wikipedia dumps from the months June and July 2015 in the languages occurring in the data, namely Greek, English, French, Polish and Amharic. Since the Amharic wikipedia is written in the Ge'ez script and the data only contains transliterated Amharic, all Amharic texts were transliterated prior to training. Then, Tweets have been segmented by assigning each word to the model with the highest probability. Training on a corpus of Twitter data, separated by language, might yield better results for the supervised approach; however, such a corpus would have to be compiled first.

For this toy example, the results show that the language model induction seems to work reasonably well with scores comparable to the supervised approach, sometimes even performing better than the supervised approach.

Closer inspection of the results reveals that the language model induction tends to generate too many clusters for a single language, resulting in a degradation of the accuracy, while on the other hand also being able to separate the different languages surprisingly well.

For example, the first tweet "Μόλις ψήφισα αυτή τη λύση Internet of Things, στο διαγωνισμό BUSINESS IT EXCELLENCE." is decomposed into two English clusters and two Greek clusters, with one erroneous inclusion of 'EXCELLENCE.' in the Greek cluster.

- Things,
- Μόλις λύση διαγωνισμό EXCELLENCE.
- Internet of BUSINESS IT
- ψήφισα αυτή τη στο

	Induction		Supervised	
	F1	F5	F1	F5
Tweet 1	0.5294	0.4775	0.7441	0.8757
Tweet 2	0.7515	**0.9325**	0.7570	0.8121
Tweet 3	0.4615	0.8185	0.6060	0.8996
Tweet 4	0.5172	0.7587	0.7360	0.9545
Tweet 5	**1.0000**	**1.0000**	0.2500	0.4642

Table 1: F-Scores

The second tweet "Demain #dhiha6 Keynote 18h @dhiparis "The collective dynamics of science-publish or perish; is it all that counts?" par David" and its decomposition. It is clear that we have one English cluster and one French cluster, and two other clusters, one of which could be labeled 'Named Entity' cluster and the other possibly 'English with erroneous inclusion of @dhiparis'. Interestingly, the French way of notating time '18h' is also included in the French cluster.

- Keynote "The collective of science-publish or perish; it all that counts?"
- Demain 18h par
- #dhiha6 David
- @dhiparis dynamics is

The third tweet "Food and breuvages in Edmonton are ready to go, just waiting for the fans #FWWC2015 #bilingualism" is split into one acronym group, three English clusters and one French cluster with the erroneous inclusion of 'go'.

- #FWWC2015
- breuvages, go
- Food, Edmonton, to, for, the
- in, waiting, #bilingualism
- and, are, ready, just, fans

The fourth tweet "my dad comes back from poland with two crates of strawberries, żubrówka and adidas jackets omg" again is split into two English clusters and one Polish cluster with the erroneous inclusion of 'back'.

- comes, from, with, two, crates, of, strawberries, jackets, omg
- my, dad, poland, and, adidas
- back, żubrówka

Finally, the last tweet "Buna dabo naw (coffee is our bread)." is decomposed as follows. The English words are split across four clusters while the transliterated Amharic text is clustered together. The splitting is due to the structure of the tweet; there is not enough overlapping information to build an English cluster.

- (coffee
- bread).
- is
- our
- Buna dabo naw

4 Conclusion

The paper has presented the early findings of a weakly supervised approach for language segmentation that works on short texts. By taking the text itself as basis for the induced language models, there is no need for training data. As the approach does not rely on external language knowledge, the approach is language independent.

The results seem promising, but the approach has to be tested on more data. Still, being able to achieve results comparable to supervised approaches with a weakly supervised method is encouraging.

5 Future work

Future work should concern the reduction of the number of generated clusters, ideally arriving at one cluster per language. Alternatively, it would be possible to smooth the frequent switching of language models by taking context into account.

Also, since the structure of the text strongly influences the presented approach, some form of text normalization could be used to increase the robustness of the system.

Sources

GaloTyri. "Μόλις ψήφισα αυτή τη λύση Internet of Things, στο διαγωνισμό BUSINESS IT EXCELLENCE.". 19 June 2015, 12:06. Tweet.

Claudine Moulin (ClaudineMoulin). "Demain #dhiha6 Keynote 18h @dhiparis "The collective dynamics of science-publish or perish; is it all that counts?" par David". 10 June 2015, 17:35. Tweet.

HBS (HBS_Tweets). "Food and breuvages in Edmonton are ready to go, just waiting for the fans #FWWC2015 #bilingualism". 6 June 2015, 23:29. Tweet.

katarzyne (wifeyriddim). "my dad comes back from poland with two crates of strawberries, żubrówka and adidas jackets omg". 8 June 2015, 08:49. Tweet.

TheCodeswitcher. "Buna dabo naw (coffee is our bread).". 9 June 2015, 02:12. Tweet.

References

Alex, Beatrice, Amit Dubey, and Frank Keller. 2007. Using Foreign Inclusion Detection to Improve Parsing Performance. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing and Computational Natural Language Learning*, pages 151–160.

Beesley, Kenneth R. 1988. Language identifier: A computer program for automatic natural-language identification of on-line text. In *Proceedings of the 29th Annual Conference of the American Translators Association*, volume 47, page 54.

Dunning, Ted. 1994. *Statistical Identification of Language*. Computing Research Laboratory, New Mexico State University.

Jain, Naman and Riyaz Ahmad Bhat. 2014. Language Identification in Code-Switching Scenario. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 87–93.

King, Ben and Steven P Abney. 2013. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 1110–1119.

Lui, Marco, Jey Han Lau, and Timothy Baldwin. 2014. Automatic detection and language identification of multilingual documents. *Transactions of the Association for Computational Linguistics*, 2:27–40.

Manning, Christopher D, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press.

Mendizabal, Iosu, Jeroni Carandell, and Daniel Horowitz. 2014. TweetSafa: Tweet language identification. *TweetLID @ SEPLN*.

Porta, Jordi. 2014. Twitter Language Identification using Rational Kernels and its potential application to Sociolinguistics. *TweetLID @ SEPLN*.

Solorio, Thamar, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the First Shared Task on Language Identification in Code-Switched Data. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing*, pages 62–72.

Yamaguchi, Hiroshi and Kumiko Tanaka-Ishii. 2012. Text segmentation by language using minimum description length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 969–978. Association for Computational Linguistics.

Zubiaga, Arkaitz, Inaki San Vicente, Pablo Gamallo, José Ramom Pichel, Inaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Vıctor Fresno. 2014. Overview of TweetLID: Tweet language identification at SEPLN 2014. *TweetLID @ SEPLN*.