Wikipedia Titles as Noun Tag Predictors

Armin Hoenen
Goethe University Frankfurt
Text Technology Lab
Robert -Mayer Str. 10
60325 Frankfurt, Germany
hoenen@em.uni-frankfurt.de

Abstract

In this paper, we investigate a covert labeling cue, namely the probability that a title (by example of the Wikipedia titles) is a noun. If this probability is very large, any list such as or comparable to the Wikipedia titles can be used as a reliable word-class (or part-of-speech tag) predictor or noun lexicon. This may be especially useful in the case of Low Resource Languages (LRL) where labeled data is lacking and putatively for Natural Language Processing (NLP) tasks such as Word Sense Disambiguation, Sentiment Analysis and Machine Translation. Profitting from the ease of digital publication on the web as opposed to print, LRL speaker communities produce resources such as Wikipedia and Wiktionary, which can be used for an assessment. We provide statistical evidence for a strong noun bias for the Wikipedia titles from 2 corpora (English, Persian) and a dictionary (Japanese) and for a typologically balanced set of 17 languages including LRLs. Additionally, we conduct a small experiment on predicting noun tags for out-of-vocabulary items in part-of-speech tagging for English.

Keywords: Wikipedia, Wiktionary, Noun, Title

1. Introduction

In linguistic literature, one finds that borrowed words are acknowledgedly overly likely to be nouns, (Haspelmath, 2008). Clues such as this one make it possible to use the information on one type of linguistic knowledge (borrowing) and to transfer it to another (word class). In this paper, we investigate, whether a large probability for being a noun likewise holds true for (single token) titles.[1] We assume, that titles as found in the Wikipedia are often definition labels, which would help explain such a bias. However, the term definition must be used with caution since Wikipedia articles are not all necessarily definitions in the strict sense. For LRLs[2] resources are lacking and state-of-the-art NLP is complicated, since supervised statistical algorithms typically get performant with large quantities of labeled data for training. Yet speaker communities do produce written resources and make use of the proliferation and input ease of texts in the internet as opposed to the difficulties and costs of publishing a printed text. Especially crowd-sourced resources such as the Wikipedia have a growing body of texts even in LRLs. The question arises of how this text can be used to improve the resource situation of an LRL. Apart from compiling a corpus, which is the most obvious use of such resources, additional layers of information can be extracted. A clue as described above (presence as Wikipedia title indicating noun as word class) could lead to additional labeled data in LRLs for which alternative sources of such labeling are unlikely to be available.

Additionally, knowing, that a token is a noun can be the foundation for other NLP tasks. For instance, in the evaluation of modern part-of-speech (POS) taggers results drop for unknown words, see for instance (Toutanova and Manning, 2000). (Toutanova and Manning, 2000) added rules based on English grammar and orthography for unknown word tagging improvement and achieved more than 15% accuracy gain. In similar postprocessing steps, noun labels could be assessed through the Wikipedia titles research.

Generally, using the following assertion, we propose a web-resource-based method to provide noun tags for any application and language: *Wikipedia single token titles have an exceedingly high probability of being nouns*. This, if it proves true, entails the possibility of using the presence of an unknown word in the titles of the Wikipedia for that language as noun tag predictor. How useful this information will be depends on the strength of the bias and on availability of other labeled data (LRL) and on the task.

2. Literature

The exploitation of collaborative platforms such as the Wikipedia is well practised in NLP, for instance in POS-tagging. In (Li et al., 2012) the authors describe how to use the Wiktionary for POS-tagging. Across the 9 languages they tested, the accuracy on unknown words was 63%.

NLP tasks such as Word Sense Disambiguation (WSD), Sentiment Analyses (SA) and Machine Translation (MT) feature research especially devoted to nouns, see for instance (Fung, 1995) or to the processing of newspaper titles, which are of course longer than Wikipedia titles, but for which a bias towards nouns could be present. Compare (Chaumartin, 2007, p.423) who finds that "a news title is

[1] Multiple token titles constitute a considerable proportion of the Wikipedia titles, differing from language to language. Their amount ranged from 41 to 68 % with a mean of 54% as approximated by using the presence of the underscore. Generally it is believed here, that single token titles are more likely to be nouns and more likely to be definitions. Apart from this, using a multiple token title as a label predictor for a single token requires to solve questions concerning overlap. The construction of this and the evaluation are not trivial and are suspended for the time being, until for the more consistent scenario, an answer has been found.

[2] There are different LRLs in terms of their dynamics. A subdivision into dead historical LRLs, dialectal LRLs, few-speaker LRLs and many-speaker LRLs may be useful. At the same time a label Medium Resource Language could be discussed.

sometimes reduced to a nominal group". Also, other resources and semantic networks, such as BabelNet, (Navigli and Ponzetto, 2012), could be of interest. For WordNet, (Navigli and Ponzetto, 2012, p.240) atest a "limited connectivity of parts of speech other than nouns in WordNet" in their setting.

3. Wikipedia and Wiktionary

Looking at the lists of all Wiktionaries and Wikipedias,[3] there is no Wiktionary, where for the same language there is no Wikipedia, while with 116 Wikipedias, the language has no Wiktionary. Comparing the sizes of the Wiktionaries and Wikipedias, a similar observation can be made, there are many more larger Wikipedias (153) than larger Wiktionaries (19). We believe, it is not by chance that this pattern occurs but reflects the dynamic of transition from print (or oral) into the digital. People are first and foremost interested in creating a knowledge database and look-up device for all kinds of real world things (mostly such things which they do not frequently use and have to look up) before they start being interested in dictionaries, structured grammar and correct orthography. In view of this, we are able to predict that our method will be especially valuable for languages where no Wiktionary is yet available and that such languages will continue to exist until all languages do have Wikipedias and Wiktionaries.

A transition of media already occurred centuries back when script was invented. (Ong, 2012) describing it, mentions that definitions as such are a literate phenomenon arising with the script mediated novel possibility to externalize knowledge and through this save knowledge about things with only marginal relevance.[4] Definitions were naturally primarily sought for things. Things typically are expressed as nouns, and "reference is primarily established through nouns", (van Hout and Muysken, 1994, p.42). Even for word families, where the typical usage of the majority of the members is not *nouny*, a noun can be derived. Thus, definitions could exclusively use noun labels but still putatively cover all domains of language. Wikipedia titles are not necessarily definitions in the strict sense, but as (Navigli and Velardi, 2010, p.1323) note: "The first sentence of Wikipedia entries is, in the large majority of cases, a definition of the page title.". If there is thus a strong bias of Wikipedia titles to be nouns, the large coincidence with definition labels should constitute an important if not the most important factor in determining such a bias.

The English Wikipedia often redirects from non-nouns including function words to nouns, for instance 'although' redirects to 'contradiction' or 'write' to 'writing'.[5] Additionally, many articles with non-noun titles such as 'forgotten' point to extremely low frequency labels such as band names and lead to disambiguation pages.

These phenomena form the theoretical foundation of a noun bias for titles in the Wikipedia. If real, how strong would that noun bias be. Could it be attested across languages, it would underscore the generalizability of the statement and enable the usage of the Wikipedia titles or more generally of definition labels as noun predictors where no labeled data is available. In fact, the phenomenon could even be more general extending to titles as mentioned above and it remains for future research to determine, whether it can be usefully applied to additional NLP tasks.

4. Language Choice

As (Rijkhoff et al., 1993) and other linguistic typologists have emphasized, when sampling languages, one should include languages of different grammatical set-ups or genealogies to avoid bias. It is not necessarily true, that an algorithm that can process a large number of similar languages is successfully handling very diverse languages. From a typological point of view, due to availability and other factors, there is a danger of overfitting to Indo-European languages written in the Latin script. However, it is very likely that statistically robust results will hold for the vast majority of language types.

For this study, we attempted to make a 30-language sample as described in (Rijkhoff et al., 1993, p.186), but ended up with 15 languages as obviously languages such as Meroitic, which is a dead historical language of Sudan, does not provide any data.[6] In choosing languages, it was tried to include languages with various Wikipedia sizes so as to have a not only typologically but at the same time a statistically balanced corpus. (Rijkhoff et al., 1993) include *Pidgins and Creoles*, but not explicitly historical or constructed languages. Thus, we added Latin and Volapük for completeness. As to the influence of the variable writing system, we included in our language sample syllabaries, logographic writing, mixed writing systems, alphabets, abjads and abugidas.[7]

Additionally, we process in greater depth three languages, whose writing systems are sufficiently different: the largely isolating Indo-European language English (Latin alphabet); the agglutinative language Japanese, written in a mix of syllabic, alphabetic and logographic characters without spaces and a rather inflectional Indo-European language written with Arabic letters, Persian.

5. Experiment 1 – Quantifying the noun bias

For Japanese, we downloaded the UniDic dictionary as used with the MeCab standard tokenizer and POS-

[3] `https://meta.wikimedia.org/wiki/List_of_Wikipedias` and `https://meta.wikimedia.org/wiki/Wiktionary#List_of_Wiktionaries` as of 21st of July 2015.

[4] Non literate peoples tend to have only words for what bares relevance to their daily lives. These things again are internalised to such a degree, that definitions are superfluous.

[5] Other languages such as German or French do not yet seem to implement that strategy with the same rigour.

[6] There was no (sufficient) data on Australian, Chukchi-Kamchatkan, Indo-Pacific, Khoisan, Sumerian, Ket, Nahali, Hurrian, Burushaki, Meroitic, Etruscan, Gilyak, Na-Dene and Nilo-Saharan languages as required by (Rijkhoff et al., 1993). Partly, these languages are dead historical languages (Etruscan, Hurrian, Meroitic, Sumerian), partly they are spoken by comparatively few speakers (e.g. Ket, Chukchi-Kamchatkan) or are language isolates (Burushaki) or a combination of those.

[7] Syllabaries write syllable characters instead of phoneme letters, logographic writing uses symbols as referents for objects or morphemes rather than sounds, abjads omit short vowels, abugidas have inherent vowels.

tagger/morphological analyzer, (Kudo et al., 2004).[8] We downloaded the titles from the Japanese Wikipedia dumps and detected the words that were present in both Wikipedia titles and the dictionary. The percentage of nouns[9] in the dictionary was 37%, whereas the percentage of nouns in the Wikipedia titles found in the dictionary was 97%. For Persian, instead of a lexicon featuring POS-tags, we used the annotated UPC corpus ((Mahmood, 2006), (Mojgan, 2012)). Again, while on token level in the corpus around 40% were nouns, the Wikipedia titles contained a much higher percentage of roughly 85%.

For English, we extracted a lexicon from the Brown Corpus, (Brown Corpus, 1979), as present in the nltk library.[10] For various comparisons, we used the mapping for the Brown Corpus provided by (Petrov et al., 2012) and for the Wiktionary, the one by (Li et al., 2012) to map the Brown Corpus tag set and the Wiktionary tag set to the Universal tag set. We then intersected the Brown Corpus lexicon with the online Wiktionary extracting for each hit the predominant POS.[11] Of 59.962 lexical entries, 17.390 remained which were not in the Wiktionary vocabulary. Excluding tokens which included non standard characters, we ended up with 12.300 unknown words. Of these 4430 were exactly found as Wikipedia titles (excluding again non standard characters as well as multi word units and deleted or redirected pages). This implies an information surplus even in case a Wiktionary does already exist for a language. Of these matches finally, 78% were nouns in the original Brown Corpus gold standard.

Hence, looking at a dictionary and two corpora, the proportion of nouns in the Wikipedia titles is much larger in all three cases than is the proportion of nouns in the source data. In order to corroborate this, we looked at the remaining 17 languages in the following way. We extracted the nouns, verbs and adjectives for the respective languages present in the English version of the Wiktionary, which has entries for almost every language.[12] This data is an approximation as other word classes are being dismissed partly due to set incompatibilities in the word class labels. However, they present the main classes of content words, which would make them most numerous as opposed to function words. It should also be noted, that the Wiktionary is a resource, which does not have an absolute character. That is it is most likely incomplete and may contain various errors or typos. However, we believe that this data is the best approximation for testing a noun bias, that can be currently used with a cross-linguistic sample.

This data itself has a probable noun bias of its own. As being part of the *English* Wiktionary, English loanwords from the respective languages should be most relevant and thus find their way into this data source quickest. For loanwords, (Haspelmath, 2008) states: "It is widely acknowledged that nouns are borrowed more easily than other parts of speech" referring to (Whitney, 1881; Moravcsik, 1978; Myers-Scotton, 2002; van Hout and Muysken, 1994).

We intersected the Wiktionary data with the Wikipedia titles (excluding multiple word tokens). The results can be seen in Table 1. Here too, for each and every language, the percentage of nouns within the matched Wikipedia titles was larger than the noun percentage of foreign words in the English Wiktionary (which covers only a small fraction of the Wikipedia titles) independent of sample size or amount of nouns in the source data. The average noun percentage in the matched Wikipedia titles was highly significantly larger than in the Wiktionary according to a one sample t-test, $t = 7.149, df = 16, p_{two-tailed} < 0.001$. Thus, it appears that single-token titles or definition labels are even more probable to be nouns than loanwords are, cross-linguistically/universally.

Language	NW	NM	Cov	M	T
Cree	0.95	1	0.0025	5	2,019
Inuktitut	0.92	0.93	0.0204	54	2,645
Nauruan	0.82	0.83	0.0057	23	4,018
Tok Pisin	0.67	0.82	0.013	71	5,473
Nahuatl	0.73	0.95	0.0021	44	20,822
Swahili	0.83	0.93	0.0162	1,186	73,091
Javanese	0.81	0.89	0.0008	92	121,633
Latin	0.45	0.88	0.0208	4,677	224,616
Volapük	0.82	1	0.0005	130	248,597
Azeri	0.9	0.96	0.0025	631	250,380
Malayalam	0.97	1	0.0005	129	262,810
Georgian	0.72	0.95	0.0151	4,081	270,823
Basque	0.87	0.93	0.0022	1,239	568,175
Hungarian	0.63	0.90	0.0103	10,664	1,031,473
Arabic	0.63	0.72	0.0037	8,572	2,289,348
Mandarin	0.64	0.81	0.0084	36,439	4,338,597
Spanish	0.63	0.82	0.007	35,775	5,127,973
Mean	0.76	**0.9**	0.008	6,106	87,3087

Table 1: NW = percentage of nouns in English Wiktionary entries (nouns, verbs, adjectives); NM = Proportion of nouns among Wikipedia titles matching the Wiktionary entries; Cov = Coverage of Wiktionary on Wikipedia titles (multi word units excluded in both cases); M = number of matches; T = number of Wikipedia titles

6. Experiment 2 – Taggers

Various tagger architectures do have different ways to handle out of vocabulary (OOV) items. (Toutanova and Manning, 2000) incorporated linguistic features on top of statistical processing. Other taggers, such as the linguaEN tagger[13] assign the label noun to all unknown words. In these cases, the Wikipedia title noun prediction is not necessarily improving performance.

[8] `http://taku910.github.io/mecab/` and `https://en.osdn.jp/projects/unidic/`

[9] 名詞

[10] http://www.nltk.org/

[11] Words especially in isolating languages can have different POS for instance depending on the syntactical position or because of homophony, compare "Can a canner can a can?". It is assumed that the Wikipedia titles characterize the main POS, that is the most frequent POS of a word as being a noun. The results on token level obtained here support this view.

[12] Example URL: `https://en.Wiktionary.org/wiki/Category:Cree_nouns`

[13] http://search.cpan.org/ acoburn/Lingua-EN-Tagger/

Proportion(Train/Test)	**90/10**	**80/20**	**70/30**	**60/40**	**50/50**	**40/60**	**30/70**	**20/80**	**10/90**
Accuracy	0.9	0.86	0.88	0.88	0.87	0.88	0.85	0.86	0.86
Coverage	0.59	0.53	0.54	0.53	0.53	0.51	0.48	0.48	0.43

Table 2: Accuracies of noun assignments through Wikipedia titles for OOV items in different training-test proportions of the Brown Corpus. Second row: Percentages of nouns found in the Wikipedia titles given all unknown tokens from the test set.

We conducted a training series of the Stanford Tagger ((Toutanova and Manning, 2000), (Toutanova et al., 2003), using the proposed basic tagger settings[14]) with the Brown Corpus for English, where we varied the file proportion between 90% and 10% for training to simulate various sizes of annotated data and amounts of unknown words. We measured how accurrately a Wikipedia title noun prediction performed. The results and the coverage can be seen in Table 2.[15] The method robustly performs well with up to 90% accuracy and the Wikipedia titles covered around half of the unknown noun instances.

Exemplarily, looking at both ends of the continuum, see Table 3, within the unknown words present as Wikipedia titles (UW) of the unknown nouns therefrom (N), the Stanford Tagger mistagged only few nouns (SMT). On the other hand, UW contained some non-nouns (NN), the majority of which led to redirects, disambiguation pages or deleted pages leaving some errors through the method (WE) but excluding relatively many assignments.

7. Conclusion and Outlook

Single token Wikipedia titles have an exceedingly large language independent probability of being nouns. This noun bias could be even stronger than the noun bias in borrowing found in linguistics. In fact, the probability of a single token title in the Wikipedia, the entirety of which can easily be downloaded being a noun is so large, that it is a promising prospect to use this information in NLP tasks. Excluding redirects, disambiguative or deleted pages improves the accuracy of Wikipedia titles as noun predictors but decreases their coverage substantially.

For determining how to improve specific NLP tasks such as current POS-taggers performance using the Wikipedia titles, more research is needed. Moreover, the method is restricted to nouns entailing if at all only very marginal gains. In the case of LRLs however, this information is disproportionately more valueable given the probable lack of labeled data. A possible application scenario would be the usage of the Wikipedia titles for unsupervised POS-tagging through clustering, where the cluster with the largest number of Wikipedia titles instances is labelled noun. The Wikipedia titles' further potential for extracting derivational or inflectional rules from redirects, the potential for extracting information on multi word units and proper nouns remain largely unexplored.

[14]The *arch* and *search* parameters were chosen as described in bullet point 11 on `http://nlp.stanford.edu/software/pos-tagger-faq.shtml`.

[15]The training and test files were randomized and each training and test set contains different data.

8. Acknowledgements

We greatfully acknowledge the support arising from the collaboration of the empirical linguistics department and computer science manifesting in the Centre for the Digital Foundation of Research in the Humanities, Social, and Educational Sciences (CEDIFOR: `https://www.cedifor.de/en/cedifor/`).

9. Bibliographical References

Chaumartin, F.-R. (2007). Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 422–425, Stroudsburg, PA, USA. Association for Computational Linguistics.

Fung, P. (1995). A pattern matching method for finding noun and proper noun translations from noisy parallel corpora. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 236–243, Stroudsburg, PA, USA. Association for Computational Linguistics.

Haspelmath, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. In Thomas Stolz, et al., editors, *Aspects of language contact: New theoretical, methodological and empirical findings with special focus on Romancisation processes*, pages 43–62. Mouton, Berlin.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). Applying Conditional Random Fields to Japanese Morphological Analysis . In Dekang Lin et al., editors, *Proceedings of EMNLP 2004*, pages 230–237, Barcelona, Spain, July. Association for Computational Linguistics.

Li, S., Graça, J. a. V., and Taskar, B. (2012). Wiki-ly supervised part-of-speech tagging. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 1389–1398, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mahmood, B. (2006). The role of the corpus in writing a grammar: An introduction to a software. *Iranian Journal of Linguistics*, 19.

Mojgan, S. (2012). *Morphosyntactic Corpora and Tools for Persian*. Ph.D. thesis, Uppsala University.

Moravcsik, E. A. (1978). Universals of language contact. In Joseph H. Greenberg, et al., editors, *Universals of human language, Volume 1, Method and theory*, pages 93–122. Stanford University Press, Stanford, CA.

Myers-Scotton, C. (2002). *Language contact: Bilingual encounters and grammatical outcomes*. Oxford University Press, Oxford.

Cond.	UW	N	SMT	NN	Redir.	Disamb.	Del.	WE
90-10	1785	1611	38	174	80	56	5	33
10-90	11734	10071	313	1663	646	677	37	303

Table 3: Wikipedia titles noun assignment details confronted with Stanford Tagger votes. Abbreviations in text.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Navigli, R. and Velardi, P. (2010). Learning word-class lattices for definition and hypernym extraction. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 1318–1327, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ong, W. J. (2012). *Orality and Literacy*. Routledge.

Petrov, S., Das, D., and McDonald, R. (2012). A universal part-of-speech tagset. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Rijkhoff, J., Bakker, D., Hengeveld, K., and Kahrel, P. (1993). A method of language sampling. *Studies in Language*, 17(1):169–203.

Toutanova, K. and Manning, C. D. (2000). Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *EMNLP/VLC 2000*, pages 63–70.

Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

van Hout, R. and Muysken, P. (1994). Modeling lexical borrowability. *Language Variation and Change*, 6:39 – 62.

Whitney, W. D. (1881). On mixture in language. *Transactions of the American Philosophical Association*, 12:1 – 26.

10. Language Resource References

Brown Corpus. (1979). *Brown Corpus Manual*. Department of Linguistics, Brown University, Providence, Rhode Island, US.