BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 14 2004, pages 2320–2321
doi:10.1093/bioinformatics/bth227

MedPost: a part-of-speech tagger for bioMedical text

L. Smith[1,], T. Rindflesch[2] and W. J. Wilbur[1]*

[1]Computational Biology Branch, National Center for Biotechnology Information and [2]Cognitive Science Branch, Lister Hill National Center for Biomedical Communications, 8600 Rockville Pike, Bethesda, MD 20894, USA

Received on December 19, 2003; revised on February 18, 2004; accepted on March 4, 2004
Advance Access publication April 8, 2004

ABSTRACT

Summary: We present a part-of-speech tagger that achieves over 97% accuracy on MEDLINE citations.

Availability: Software, documentation and a corpus of 5700 manually tagged sentences are available at ftp://ftp.ncbi.nlm.nih.gov/pub/lsmith/MedPost/medpost.tar.gz

Contact: lsmith@ncbi.nlm.nih.gov

MEDLINE is a bibliographic database of publications in health sciences, biology and related fields. It currently contains over 12 million records, and nearly 7 million include an abstract. The NCBI PubMed website[1] provides an interface for searching MEDLINE and retrieving documents in several different formats. There is a growing amount of natural language processing research using biomedical text, especially MEDLINE abstracts, to improve access to the literature (information retrieval), to build databases of knowledge (information extraction) and to support automated reasoning (knowledge discovery). This research requires increasingly effective computer comprehension of language, the starting-point for which is part-of-speech tagging, or determining the syntactic function of words in text.

The value of part-of-speech tagging degrades rapidly as the error rate increases. For example, an error rate as low as 4% corresponds approximately to one error per sentence, which may severely limit the number of sentences that can be analyzed successfully. Taggers developed for general text do not perform well when applied to MEDLINE. For example, the Brill tagger (Brill, 1992) applied to 1000 sentences, selected randomly from MEDLINE, achieved an accuracy of 86.8% using the Penn treebank tag set. This poor performance may be due to the specialized vocabulary of MEDLINE. For example, we found that nearly 57.8% of the token types in MEDLINE did not occur in either the Brown corpus (Marcus *et al*., 1994) or the AP corpus (1988/1989 version) (based on 92.7% of the most common tokens in each corpus).

Our tagger (called MedPost) was developed to meet the need for a high accuracy part-of-speech tagger trained on the MEDLINE corpus. On the 1000 sentence test set, it achieves an accuracy of 97.43% using its native tag set and 96.9% accuracy using the Penn treebank tag set. Approximate 95% confidence intervals for these figures are within ± 0.25.

The medpost program can be run on most Unix operating systems with standard utilities (gunzip, tar, make, gcc, perl, nroff). Instructions for installing the program are contained in the file INSTALL.medpost, which can be found in the distribution, and details on running the program can be found in a *man* page, which is provided.

The program currently accepts text for tagging in either native MEDLINE format or XML, both available as save options in PubMed. In addition, it recognizes a simpler 'ITAME' format that allows text (with optional title and identifier) from any source to be tagged. The tagger segments input text into sentences and output each token with a part-of-speech tag separated by an underscore. For example, this is the result of tagging sentence number 9 from the MEDLINE abstract with PMID 1847596,

> Surprisingly_RR ,_, NO3-_NN inhibited_VVD
> the_DD rate_NN of_II K+_NN
> swelling_VVGN by_II 82_MC %_SYM ._.

A command line option directs the tagger to translate the output to either the Penn treebank (Marcus *et al*., 1994) or SPECIALIST lexicon tag set (National Library of Medicine, 2003). Here is the same sentence after translation to the Penn treebank tag set,

> Surprisingly/RB ,/, NO3-/NN inhibited/VBD
> the/DT rate/NN of/IN K+/NN swelling/NN by/IN
> 82/CD %/SYM ./.

The MedPost tag set consists of 60 part-of-speech tags listed in Table 1. It was derived from the Penn treebank tag set (Marcus *et al*., 1994), a subset of the UCREL tag set (Garside *et al*., 1997), and a generalization of the SPECIALIST lexicon tag set (National Library of Medicine, 2003). Our goal was

*To whom correspondence should be addressed.

[1]see http://ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed

Table 1. The MedPost part-of-speech tag set

CC	coordinating conjunction (17/991)	RR	adverb (17/651)	VHG	participle *having* (0/0)
CS	subordinating conjunction (10/120)	RRR	comparative adverb (1/1)	VHI	infinitive *have* (0/5)
CSN	comparative conjunction (than) (2/56)	RRT	superlative adverb (1/14)	VHN	participle *had* (0/0)
CST	complementizer (that) (3/122)	SYM	symbol (0/289)	VHZ	3rd pers. sing. *has* (0/28)
DB	predeterminer (0/7)	TO	infinitive marker *to* (3/159)	VVB	base form lexical verb (21/209)
DD	determiner (25/2089)	VM	modal (1/112)	VVD	past tense (64/306)
EX	existential *there* (2/19)	VBB	base *be, am, are* (1/147)	VVG	present part. (15/144)
GE	genitive marker *'s* (0/12)	VBD	past *was, were* (0/453)	VVI	infinitive lexical verb (9/127)
II	preposition (27/3470)	VBG	participle *being* (0/5)	VVN	past part. (60/815)
JJ	adjective (64/2302)	VBI	infinitive *be* (0/35)	VVZ	3rd pers. sing. (7/133)
JJR	comparative adjective (3/63)	VBN	participle *been* (0/31)	VVNJ	prenominal past part. (32/322)
JJT	superlative adjective (0/13)	VBZ	3rd pers. sing. *is* (0/162)	VVGJ	prenominal present part. (21/135)
MC	number or numeric (21/970)	VDB	base *do* (0/4)	VVGN	nominal gerund (44/152)
NN	noun (97/6344)	VDD	past *did* (0/16)	(	left parenthesis (0/456)
NNP	proper noun (18/30)	VDG	participle *doing* (0/0)	)	right parenthesis (0/463)
NNS	plural noun (42/2014)	VDI	infinite *do* (0/0)	,	comma (0/963)
PN	pronoun (3/124)	VDN	participle *done* (0/1)	.	end-of-sentence period (0/1000)
PND	determiner as pronoun (29/66)	VDZ	3rd pers. sing. *does* (0/5)	:	dashes, colons (0/115)
PNG	genitive pronoun (0/89)	VHB	base *have* (5/40)	“	left quote (5/10)
PNR	relative pronoun (7/126)	VHD	past *had* (0/45)	”	right quote (4/13)

The number of errors per number of occurrences is given for each tag in the 1000 sentences of the test set. Overall, the tagger achieved an accuracy of 97.43% on 26 566 tokens, with 582 sentences tagged without any errors and 261 tagged with a single tagging error.

to make the tags as unambiguous as possible, to limit their number, and to enable easy and unambiguous translation to the Penn treebank and SPECIALIST lexicon tag sets.

To test and train the tagger, 5700 sentences (155 980 tokens) were selected randomly from various thematic subsets (Wilbur, 2002) of MEDLINE, and were manually tagged. The criteria for deciding membership in word classes were Quirk *et al*. (2000); Summers (2003).

Processing begins with a perl script of regular expressions that tokenizes the input following the conventions of the Penn treebank (Marcus *et al*., 1994) and that locates sentence boundaries (usually periods, except for decimal points and abbreviations). The tokens of each sentence are then passed to a stochastic tagger that employs a hidden Markov model (HMM) (Rabiner, 1988). Each part-of-speech tag corresponds to a state in the model, and transition probabilities are estimated from tag bigram frequencies in the training set. The output probabilities of the HMM are determined for words in the lexicon assuming equal probability for the possible tags. Output probabilities for unknown words are based on word orthography (e.g. upper or lower case, numerics, etc.), and word endings up to four letters long. The Viterbi algorithm is used to find the most likely tag sequence in the HMM matching the tokens.

We found that high-accuracy tagging required a high coverage lexicon for ‘open class’ words (nouns, verbs, etc.). Therefore, a lexicon of 10 000 open class words was created for the most frequently occurring words in MEDLINE (accounting for 92.7% of its tokens). In addition, all ‘closed class’ words (determiners, pronouns, etc.) were included in the lexicon. The entry for each word in the lexicon includes a manually entered list of the allowed part-of-speech tags. For a small proportion of words, and for word endings of unknown words, the entry also specifies a priori probabilities for the tags. But for most words, the allowed tags are assumed to occur with equal probability. Despite the lack of probability information for most words, the tagger is able to achieve high accuracy by using the contextual information in the HMM to resolve ambiguities.

REFERENCES

Brill,E. (1992) A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, Trento, Italy.

Garside,R., Leech,G. and McEnery,A. (1997) *Corpus Annotation*. Longman, London and New York.

Marcus,M.P., Santorini,B. and Marcinkiewicz,M.A. (1994) Building a large annotated corpus of English: the Penn Treebank. *Computat. Linguist*., **19**, 313–330.

National Library of Medicine (2003) *UMLS Knowledge Sources*, 14th Edn.

Quirk,R., Greenbaum,S., Leech,G. and Svartvik,J. (2000) *A Comprehensive Grammar of the English Language*. Longman, London and New York.

Rabiner,L.W. (1988) A tutorial on hidden markov models and selected applications in speech recognition. *Proc. IEEE*, **77**, 257–286.

Summers,D. (eds) (2003) *Dictionary of Contemporary English*. Longman, 4th edn, London and New York.

Wilbur,W.J. (2002) A thematic analysis of the AIDS literature. *Pac. Symp. Biocomput.*, **7**, 386–397.