Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2010-012 February 25, 2010

Performance and error analysis of three part of speech taggers on health texts

Dorothy Curtis and Qing Zeng

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Performance and error analysis of three part of speech taggers on health texts

Dorothy Curtis[1], Qing Zeng [2]

[1]Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, Massachusetts, USA

[2] Department of Biomedical Informatics, University of Utah , School of Medicine, Salt Lake City, Utah, USA

Email addresses:

DC: dcurtis@csail.mit.edu

QZ: q.t.zeng@utah.edu

Abstract

Increasingly, natural language processing (NLP) techniques are being developed and utilized in a variety of biomedical domains. Part of speech tagging is a critical step in many NLP applications. Currently, we are developing a NLP tool for text simplification. As part of this effort, we set off to evaluate several part of speech (POS) taggers. We selected 120 sentences (2375 tokens) from a corpus of six types of diabetes-related health texts and asked human reviewers to tag each word in these sentences to create a "Gold Standard." We then tested each of the three POS taggers against the "Gold Standard." One tagger (dTagger) had been trained on health texts and the other two (MaxEnt and Curran & Clark) were trained on general news articles. We analyzed the errors and placed them into five categories: systematic, close, subtle, difficult source, and other. The three taggers have relatively similar rates of success: dTagger, MaxEnt, and Curran & Clark had 87%, 89% and 90% agreement with the gold standard, respectively. These rates of success are lower than published rates for these taggers. This is probably due to our testing them on a corpus that differs significantly from their training corpora. The taggers made different errors: the dTagger, which had been trained on a set of medical texts (MedPost), made fewer errors on medical terms than MaxEnt and Curran & Clark. The latter two taggers performed better on non-medical terms and we found the difference between their performance and that of dTagger was statistically significant. Our findings suggest that the three POS taggers have similar correct tagging rates, though they differ in the types of errors they make. For the task of text simplification, we are inclined to perform additional training of the Curran & Clark tagger with the Medpost corpus because both the fine grained tagging provided by this tool and the correct recognition of medical terms are equally important.

Introduction and Background

Much medical information exists as free-form text --- from patient histories, through discharge summaries, to journal articles detailing new discoveries and information about participation in clinical trials. Extracting information from this free-form text is important to both patients and caregivers. Patients want to understand their diagnosis and care instructions. Caregivers and their institutions want to ensure that appropriate care and follow-up are provided to patients. Natural Language Processing (NLP) techniques can be used to extract this information automatically. For instance, discharge summaries can be analyzed for prescriptions and dosage information appropriate to the patient's diagnosis. [1] Imaging study summaries can be analyzed to make sure that follow-up is done for patients who have new or expanding neoplasms. [2] Other applications include information extraction, text summarization, data mining, and text simplification and explanation. The MetaMap Transfer Project (MMTx) [3, 4], an NLP subsystem, has been used for many such applications, such as extracting problem lists from free-text clinical documents. [5]

Part of speech (POS) taggers are widely used in NLP applications, as an early step in the analysis of the text, to annotate each word with its part of speech. POS taggers are faster and simpler than parsers and thus more appropriate for applications where full sentence structure is not needed. This specific study was directly motivated by a consumer health informatics project on text simplification. Many studies have shown that the readability of the health information provided to consumers does not match their reading levels [6]. Despite the efforts of healthcare providers and writers to make more readable materials, most patient-oriented Web sites, pamphlets, drug-labels, and discharge instructions still require consumers to have a tenth grade reading

level or higher [7]. Concerning the extent of limited health literacy, Paasche-Orlow et al [8] pooled 85 studies of adult health literacy in the U.S. and found that 26% of the people surveyed had "low health literacy" (6^{th} grade or below) and another 20% had "limited health literacy" (7^{th} or 8^{th} grade). Note that most of these studies excluded adults who did not speak English.

To address this problem, we have proposed the development of computer-based methods for providing consumers with texts of appropriate readability levels. Our methods involve translating complex health texts into target readability levels using NLP techniques with no loss of critical information. As an initial step, we evaluated POS taggers that can be potentially used in our project.

A POS tagger, or tagger, is a software program that takes sentences as input and produces output that associates a POS tag with each input word or phrase. Prior to tagging, there are typically earlier processing stages that extract sentences from texts and handle punctuation. Taggers do not look beyond the current sentence in their work. Taggers are not parsers in that, for the most part, taggers tag individual words, while parsers try to discern the components of the sentence and how they relate to each other. When taggers tag multi-word units, typically, they are finding noun phrases or units where some of the parts cannot have individual tags. An example of this is "diabetes mellitus." "Mellitus" can, at best, be tagged individually as a "Foreign word." It has no English meaning except in the context of the word "diabetes."

One difficulty in tagging English words comes from the fact that written English is derived from spoken English and consequently some phrases have multiple written versions. Consider the example of hyphenated, or compound, words, such as "ankle-fracture": When written as "ankle fracture," it has the same meaning and plays the same role in a sentence. Similarly there are issues with spelling variations and the treatment of punctuation and other special characters.

The word "ankle-fracture" is also an example of a "constructed" word: one category of "constructed" words in English comes from gluing words together with hyphens. For taggers that depend on lexicons, i.e., lists of words and possible meanings, constructed words are a challenge, as most of them will not occur in the lexicon. Similarly, proper nouns will not be in the lexicon. In medical texts, drug names are another source of potentially unknown words. The management of unknown words is one of the greatest challenges in POS tagging.

Another difficulty that taggers face is ambiguity. Even if a word occurs in a lexicon, it may have many senses or meanings. A common example from the medical domain is "dose." "Dose" can be a noun, meaning the amount of medicine the patient should take, or it can be a verb, meaning the activity of giving medication to a patient

Taggers vary in how they are constructed. Some are rule based, while others are "statistical." The statistical taggers are "trained" on a corpus, i.e., they "learn" to tag sentences. There is little control on how well they will perform on texts outside the "area" of their training corpus. For instance, many taggers are trained on news

articles. How well they will do on texts related to health topics is unclear. Thus it is important to evaluate them for our purposes.

Taggers also vary in the sets of tags that they use. Detailed, or fine-grained taggers annotate whether a noun is singular or plural, common or proper, whether an adjective or adverb is comparative or superlative, and give some information about verbs, such as number, person, tense or other form. Less detailed, or coarse-grained taggers typically use the tag "noun" or "verb" and supply no further information. Van Halteren [9] presents more details on how taggers can differ and the challenges involved in comparing them.

In this study, we chose to examine three taggers: one trained on medical abstracts and two maximum entropy taggers trained on the Penn Treebank Wall St. Journal corpus [10]. This allowed us to assess the effect of training on medical texts versus training on a standard corpus on tagger performance and whether the two maximum entropy taggers differed from each other. dTagger [11] V0.0.2 was created by Guy Divita. It is a statistical parser that uses a Hidden Markov Model and the Viterbi algorithm. It was trained on health texts. The performance of this tagger is reported as 95% on a modified version of the MedPost hand-annotated Medline abstracts [12]. MXPOST is a maximum entropy tagger [13] created by Advaith Ratnaparkhi. The MXPOST tagger is a statistically trained tagger with additional rules for using context to improve tagging. It was trained on the Penn Treebank Wall St. Journal corpus. It claims 96.3% accuracy on unseen text. The Curran & Clark maximum entropy tagger [14] was provided by Claire Grover as part of a pre-release version of LT-TTT2 [15]. Like MXPOST, the Curran & Clark tagger is trained on the Penn Treebank data. It

has some rules-based components and some other components that use maximum entropy modelling. It claims about 97% accuracy.

This paper presents our approach to comparing POS taggers. First we describe our method for evaluating a tagger. Next, we report our results in comparing three taggers. Finally, we present our conclusions.

Methods

We tested the taggers on a data set selected from the Health Information Readability Corpus [16]. This corpus consists of 351 documents concerning diabetes, collected from the web, in six categories: consumer health information; information targeted at children; news; journals; electronic medical records; and clinical trials. We randomly chose twenty sentences from each of the six categories, for a total of 120 sentences (2375 tokens).

Gold Standard

To evaluate the taggers on these selections, we needed to create a reference point that contained "correct" tags, i.e., tags against which the taggers's tags could be compared. We called this reference point "the gold standard." Our strategy was to have a fine-grained labelling system to collect information from human taggers. For the gold standard tagging effort, we chose a tag set, roughly based on the "Part-of-Speech Tagging Guidelines for the Penn Treebank Project" [17]. In addition to the usual noun and verb tags, we added "Group-left" tags so that phrases, such as "blood pressure," could be identified as a single unit. Table 1 lists our tags, the Treebank tags, which are used by the MaxEnt and Curran & Clark taggers, and the dTagger tags.

Table 1 - Tagsets

Gold TagSet	Penn Treebank Tagset	dTagger Tagset
Determiner	DT	Det
Adjective	JJ	Adj
Adjective-comparative	JJR	adj
Adjective-superlative	JJS	adj
Adverb	RB	adv
Adverb-comparative	RBR	adv
Adverb-superlative	RBS	adv
Predeterminer	PDT	
Noun-singular	NN	noun
Noun-plural	NNS	noun
Noun-proper	NNP	noun
Noun-proper-plural	NNPS	noun
Possessive	POS	pos
Pronoun	PRP	pron

Possessive-pronoun	PRP$	pron
Acronym		
Verb-present-tense-3rd-singular	VBZ	verb
Verb-present-tense-not-3rd-singular	VBP	verb
Verb-past-tense	VBD	verb
Verb-base-form	VB	verb
Verb-gerund	VBG	
Verb-past-participle	VBN	
Modal	MD	modal
Coord-conj	CC	conj
Preposition-Subord-Conj	IN	prep/comp
Particle	RP	
To-Infinitive		
Cardinal-number	CD	

Foreign-word	FW	
Interjection	UH	
Other		
Group-left		

The gold standard tagging process for the first data selection started with two human taggers independently tagging a first batch of eighteen sentences containing 395 tokens. These taggers are computer scientists with an interest in linguistics. The tagging process was web based: each word was presented, highlighted within its sentence, on a web page. This page included a link to an online dictionary definition for the word, a link to a page of brief definitions of the tags in the gold tag set, a link to the Penn Treebank Tagging Guidelines, a set of check boxes corresponding to the gold tags and a space for comments. The human taggers agreed on 316 words. The remaining words were presented to and discussed with two other people, including a professional linguist. This team was able to reach agreement on 360 words, some of which had dual tags, i.e., either of two tags is acceptable for certain words. For 35 words, there was no agreement and were omitted from the gold standard. "Ankle-fracture" is an example of a word where there was no agreement: some human taggers considered it a "noun," while others considered it an "adjective," in the phrase "…on ankle-fracture patients…" Table 2 shows a sentence as tagged by the human taggers. One possible approach is to use a tagset with finer granularity. Such a tagset could contain a category "Noun used as adjective" to cover this situation. While there is benefit to ever finer granularity of tags in that each word can be tagged more

accurately, the merits of very fine granularity of tagging in comparing software taggers are unclear.

Table 2 - A sentence as tagged by the human taggers

Word	Tag
It	Pronoun
takes	Verb-present-tense-3rd-singular
commitment	Noun-singular
to	To-Infinitive
change	Verb-base-form
your	Possessive-pronoun
habits	Noun-plural
To	To-Infinitive
Include	Verb-base-form
Exercise	Noun-singular

The tagging principles for the first data set were established through consensus and discussion. These tagging principles were then applied to a second batch of 102 sentences. A single human tagger tagged these 102 sentences. No reliability testing was performed. Though, when the taggers disagreed with the gold standard, the gold tag was reconsidered and, in some cases, corrected.

To accommodate the varying degrees of granularity among the different tag sets, we wrote a program to compare the tags assigned to the words in our test set. This

comparator has maps from gold tags to the less detailed tags that some taggers use, e.g., it maps “Verb-past-tense” to “verb.”

Results

Table 3 shows that the three taggers have about the same rate of success: about 87%-90% of the tagged items agree with the gold standard. The total number of tagged items is different for the three taggers because the MaxEnt and Curran & Clark taggers tag single words, while dTagger tags multiple word units. Using the Chi-square test for independence, we get a test statistic of 6.19 when comparing the three taggers. This exceeds the critical value $\chi^2_{.05;2} = 5.99$, so we can conclude that there is a significant difference in the performance of these taggers. Using the same χ^2 test for independence, we get a test statistic of 1.32 when comparing the MaxEnt and Curran & Clark taggers. This does not exceed the critical value $\chi^2_{.05;1} = 3.84$, so we can conclude that there is no significant difference in the performance of the MaxEnt and Curran & Clark taggers.

Table 3 - Performance of the three taggers

	MaxEnt	dTag	Curran & Clark
Correct	2094 / 88.6%	1839 / 87.3%	2073 / 87%
Incorrect	269 / 11.4%	268 / 12.7%	239 / 13%
Total	2363 / 100%	2107 / 100%	2312 / 100%

Table 4 shows the types of errors that the three taggers made. The following subsections discuss the different types of errors.

Table 4 - Classification of tagger errors

Problem	MaxEnt	dTagger	Curran & Clark
Systematic, correctable error	4	47	6
Close error	112	84	100
Other	113 (34 medical)	109 (14 medical)	97 (24 medical)
Subtle errors	25	24	24
Difficult source	8	6	9
Total	262	270	236

Systematic Errors

DTagger made some "systematic errors" in tagging. By "systematic error," we mean that, given a particular word in a particular context, the tagger will always tag it incorrectly. This behaviour is attributed to tagging errors in the corpus on which dTagger was trained. The instances of this error that we observed in dTagger's labelling concerned the verbs "have" and "be." "Have" has two different uses in English. As a transitive verb, it indicates ownership, as in "I have a book." As a verbal auxiliary it is used to form the present perfect, past perfect, or future perfect. An example is "I have read this article many times." Here "have" helps indicate the tense of the verb "read" and has nothing to do with "ownership." When "have" occurs in a context where it indicates "ownership," it should be labelled as a "verb" and not as an "auxiliary." dTagger always tags "have" as an auxiliary. Similarly for the verb "be." Labelling "is" as "aux" could confuse a parser into believing that a sentence is "passive" when it isn't. This is important to us because a standard strategy

for improving the readability of texts is to rewrite sentences from the passive voice to the active voice. [18]

Close Errors

With "close" errors, the tagger is on the right track, but due to disagreements among the tagsets, it seems inappropriate to count the tag as correct. One example of an error that is classified as close is the labelling of "if" as a "conjunction" by dTagger. The other software taggers and the human taggers labelled "if" as a "Preposition/Subordinating Conjunction."

Subtle Errors

Similar to "close" errors, subtle errors are due to some fixed rules within a tagger that are at odds with the Gold Standard. In these cases, typically English grammar manuals are also inconsistent. One word in this category is "each," which the Gold Taggers and many online sources labelled as an "Adjective", but the Penn Treebank conventions insist on labelling it as a "Determiner." Another difficult word is "everyone." Here the Penn Treebank conventions label "everyone" as a "noun" and the Gold Taggers went along with this, but the dTagger and most sources consider "everyone" a "pronoun." A third word is "all" in the sentence, "Buy all your insulin from one pharmacy." Here the Gold Tagger label was "pronoun," while the taggers declared it to be a "noun." We believe these conflicts may not seriously affect our planned rewriting rules.

Difficult Source

Another source of error for taggers is that people can invent usages of words "on the fly" that are completely understandable to other people, but are very unusual. (This is

a different concept from a word that a tagger sees rarely.) The example that occurred in our data is the phrase "A 66-year-old gentle presenting with …" The "gold standard" team consistently labelled "gentle" as a "noun." Our opinion is that "gentle" is short-hand for "gentleman," and not the use of an adjective as a noun.

Other

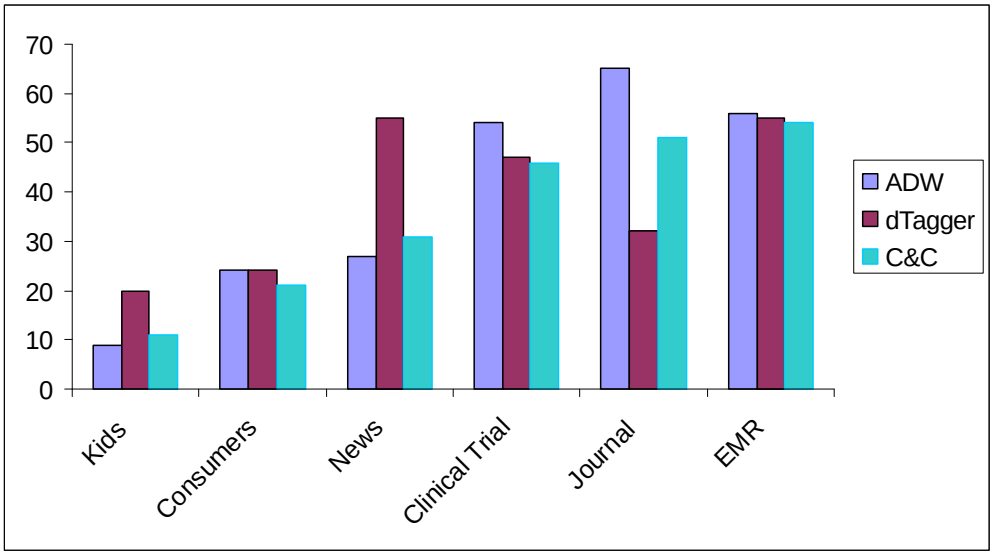
Some tags seem to be incorrect without any mitigating circumstances: labelling "exercise" as a "verb" in the sentence, "Try these strategies for adding more exercise to your life," is wrong. Similarly, labelling "try" in the same sentence as a "noun," does not seem reasonable. Labelling "routine" in the phrase "routine follow up" as a "noun" is similarly wrong. One can understand that probably these words weren't used in these contexts in the training sets on which the taggers were trained. Another case is labelling "renal" as a "noun."

While this particular classification of errors may be imprecise or suboptimal, the process of classifying these errors has informed of us of the degrees of disagreement that parts of speech taggers can have: while one is tempted to claim that the differences are small and are not worth resolving, to be able to use these programs to manipulate English texts, one needs to understand these details. At the other end of the spectrum, labelling "diabetes" or "electrolytes" or "polyps" as a "verb" or "DIABETES" as a "pronoun" or "Kidney" as a "verb" is not helpful at all. Besides the word itself being mislabelled, the parsing of the rest of the clause and sentence is severely hindered. How to deal with these sorts of errors is an open research question.

Figure 1 shows the performance of the taggers across the different types of articles in our corpus. dTagger made fewer errors than the other taggers on sentences from

journal articles, presumably because it was trained on medical abstracts. Curran & Clark and MaxEnt made fewer errors than dTagger on news articles, again, probably because they were trained on news articles.

Figure 1 – Distribution of tagger errors by article type



Discussion

We evaluated three POS taggers: the dTagger that was trained on health texts and uses a medical lexicon and the MaxEnt and Curran & Clark taggers that were trained on general news articles. Overall, the taggers achieved roughly the same rate of success (87%-90%). The errors they made did differ, however, reflecting the different training sets and the different lexicons that they used. For instance, few of dTagger's errors appeared to be caused by medical terminology.

All three taggers performed worse in our evaluation than on their original training/testing sample, the accuracy rates dropping from mid to upper 90% to upper 80%. The lower accuracy rates should be noted since many biomedical NLP applications employ POS taggers as part of the processing pipelines.

Prior research has suggested the benefit of training on domain-specific text corpus [19]. The fact that dTagger made few medical term-related errors is consistent with the prior findings. However, there are significant differences among different biomedical text types. The MedPost corpus consists of Medline abstracts and lacks five out of six types of documents that were used in this study (consumer health information, information targeted at children, news, journals, medical records, and clinical trials). We believe this contributed to the lower performance of dTagger on our test data (87%). Considering MaxEnt and Curran & Clark were not trained on a medical corpus, they both performed reasonably well (89% and 90%, respectively). They also could benefit from the additional training on medical texts and on document types other than the news (e.g. medical records).

Though we did not evaluate the taggers in the context of a specific application, taggers are used in a context. They are positioned to handle some low-level well-defined work so that writers of "higher-level" tools can focus on their desired analysis. One such usage is vocabulary simplification. In this context, it is important to identify nouns and if the tagger mislabels verb forms this has no effect on the task at hand. Sometimes taggers are used on text that is not well-formed, i.e., not grammatical. It is important to handle this case well, if the application is concerned with medical records. These texts often omit the subject from sentences. Further,

they often have many ambiguous abbreviations. In other contexts, such as syntactic transformation, the tagger's output feeds into a parser. Here poor tagger performance can confuse or mislead the parser and the errors can carry over. Given that each tagger has its strengths and weaknesses, Van Halteren [9] suggests that multiple taggers and possibly a tagger combined with a parser can yield better results than a tagger by itself.

For our research on health text simplification that involves both vocabulary and syntax, we are inclined to retrain MaxEnt or Curran & Clark with the Medpost corpus because the fine grained tagging of MaxEnt and Curran & Clark is needed by parsers and the correct recognition of medical terms is also necessary to process our health texts appropriately. Beyond that we plan to investigate the use of "Heuristic Sample Selection" as described in [20] to extend the training set as necessary.

Conclusions

We evaluated three POS taggers: the dTagger that was trained on health texts and uses a medical lexicon and the MXPOST and Curran & Clark taggers that were trained on general news articles. Overall, the taggers achieved roughly the same rate of success (87%-90%). The errors they made did differ, however, reflecting the different training sets and the different lexicons that they used. For instance, few of dTagger's errors appeared to be caused by medical terminology. The taggers' different types of errors have implications for applications that use them.

Acknowledgements

This work was supported by NIDDK grant number DK075837. We would like to thank Freddy Bafuka for his tagging efforts and Allen Browne and Guy Divita of the National Library of Medicine for their discussions of taggers and their tagging efforts. We would also like to think Carlos Nakamura for his assistance in the statistical analysis of the taggers and Carlos Nakamura and Craig Schaffert for their review of the manuscript.

References

1. Ertle AR, Camplbell EM, Hersh WR. **Automated application of clinical practice guidelines for asthma management.** In *AMIA Annual Symposium Proceeding,* 1996: 552-6.
2. Zingmond D, Lenert LA. **Monitoring free-text data using medical language processing.** *Computers and Biomedical Research* 1993 **26**(5):467-81.
3. MetaMap MMTx http://mmtx.nlm.nih.gov/
4. Bashyam V, Divita G, Bennett DB, Browne AC, Taira RK: **A normalized lexical lookup approach to identifying UMLS concepts in free text.** *Medinfo* 2007, **12**(Pt 1):545-9
5. Meystre s, Haug PJ. **Evaluation of Medical Problem Extraction from Electronic Clinical Documents Using MetaMap Transfer (MMTx).** *Studies in Health Technology and Informatics* 2005, **116**:823-8.
6. Rudd RE, Moeykens BA, Colton TC: **Health and literacy: a review of medical and public health literature**. *Annual Review of Adult Learning and Literacy* 2000:**1**:158-199
7. Nielsen-Bohlman L, Panzer AM, Kindig DA: **Health literacy: a prescription to end confusion.** Institute of Medicine. 2004.
8. Paasche-Orlow MK, Parker RM, Gazmararian JA, Nielsen-Bohlman LT, Rudd RR: **The Prevalence of Limited Health Literacy** *Journal of General Internal Medicine* 2005 **20:2**: 175–184.
9. van Halteren H: *Syntactic Wordclass Tagging*. Kluwer Academic Publishing, the Netherlands, 1999.
10. Marcus MP, Santorini B, Marcinkiewicz MA: **Building a large annotated corpus of English: the Penn Treebank.** In *Computational Linguistics,* 1993:**19:2**:313-330
11. dTagger. Available from http://lexsrv3.nlm.nih.gov/SPECIALIST/Projects/dTagger/current/index.html
12. Divita G, Browne AC, Loane R: **dTagger: a POS tagger**. In *AMIA Annual Symposium Proceedings*, 2006;:200-3
13. Ratnaparkhi A. **Maximum entropy model for part-of-speech tagging.** In **Conference on Empirical Methods in Natural Language Processing Proceedings.** *Association for Computational Linguistics,* 1996::133—142
14. Curran, J. R. and S. Clark (2003). **Investigating GIS and smoothing for maximum entropy taggers.** In *Proceedings of the 11th Meeting of the European Chapter of the Association for Computational Linguistics* (EACL-03), pp. 91–98.

15. Grover C, Matheson C, Mikheev A, Moens M: **LT TTT – A Flexible Tokenisation Tool.** In *Proceeding of the Second International Conference on Language Resources and Evaluation (LRECC 2000),*: available from http://www.ltg.ed.ac.uk/software/lt-ttt2

16. Kandula S, Zeng-Treitler Q: **Creating a Gold Standard for the Readability Measurement of Health Texts**. In *AMIA Annual Symposium Proceedings*, 2008:353-7.

17. Santorini B. **Part-of-Speech Tagging Guidelines for the Penn Treebank Project** (3rd Revision, 2nd printing). June 1990: Available at http://www.cis.upenn.edu/~treebank/home.html and ftp://ftp.cis.upenn.edu/pub/treebank/doc/tagguide.ps.gz

18. Plain Language http://www.plainlanguage.gov/

19. Codena AR, Pakhomovb SV, Andoa RK, Duffy PH, Chute CG: **Domain-specific Language Models and Lexicons for Tagging.** *Journal of Biomedical Informatics* 2005:**38:6**:422-430.

20. Liu K, Chapman W, Hwa R, Crowley RS: **Heuristic Sample Selection to Minimize Reference Standard Training Set for Part-Of-Speech Tagger**. *JAMIA* 2007: **(14:5)**:641-50.