ORIGINAL RESEARCH

Leveraging Psycholinguistic Resources and Emotional Sequence Models for Suicide Note Emotion Annotation

Eric Yeh[1], William Jarrold[2] and Joshua Jordan[2]

[1]SRI International, Menlo Park, CA. [2]University of California at Davis, Davis, CA. Corresponding author email: yeh@ai.sri.com; wljarrold@ucdavis.edu; jtjordan@ucdavis.edu

Abstract: We describe the submission entered by SRI International and UC Davis for the I2B2 NLP Challenge Track 2. Our system is based on a machine learning approach and employs a combination of lexical, syntactic, and psycholinguistic features. In addition, we model the sequence and locations of occurrence of emotions found in the notes. We discuss the effect of these features on the emotion annotation task, as well as the nature of the notes themselves. We also explore the use of bootstrapping to help account for what appeared to be annotator fatigue in the data. We conclude a discussion of future avenues for improving the approach for this task, and also discuss how annotations at the word span level may be more appropriate for this task than annotations at the sentence level.

Keywords: emotion detection, natural language processing, suicide note, psycholinguistic resources

This article is available from http://www.la-press.com.

Introduction

We describe the joint submission entered by SRI International and University of California at Davis for track 2 of the 2011 Medical NLP Challenge.[1] Our system implements a machine learning approach, and leverages a set of psycholinguistic resources to capture the emotional content of the text.

System overview

We leverage a machine learning based model for our system, using logistic regression combined with L2 regularization. Given a note, our system treats each of that note's constituent sentences as individual instances to be featurized. During training, we primarily consider each labeled instance individually. During test time, for a given note we process its sentences in sequential order, recording the annotations made for each sentence.

Our system consists of two stages, the first stage determines whether a given sentence contains any emotion annotations, the second determines which emotions should be present. Our choice of a two stage architecture was governed by the highly skewed statistics found in the training notes. As seen in Table 1, which lists the distribution of the number of emotion annotations per sentence in the training notes, the majority of sentences do not have any annotations. Table 2 lists the distribution of emotion annotations found in the training set, in descending order of frequency. If we consider the lack of any annotations for a sentence as its own distinct *no emotion* annotation, we would find that the additional 2460 *no emotion* labels would significantly outnumber all of the other annotated emotions. Given logistic regression can be very sensitive to class imbalances in the training data, this could very well result in a system with poor recall over the other emotion annotations.

Table 1. Distribution of the number of emotion annotations observed per sentence.

No. annotations	Sentences observed	Percentage
0	2460	53%
1	1871	40%
2	266	5%
3	27	0%
4	7	0%
5	2	0%

Table 2. Emotion annotations, frequency and overall percentage.

Emotion	Times observed	Percentage
Instructions	820.0	32%
Hopelessness	455.0	18%
Love	296.0	11%
Information	295.0	11%
Guilt	208.0	8%
Blame	107.0	4%
Thankfulness	94.0	3%
Anger	69.0	2%
Sorrow	51.0	2%
Hopefulness	47.0	1%
Fear	25.0	0%
Happiness/peacefulness	25.0	0%
Pride	15.0	0%
Abuse	9.0	0%
Forgiveness	6.0	0%

To prevent our system from skewing in favor of not emitting any emotion annotations, the first stage of our system performs a binary classification, identifying whether a sentence should have any emotions annotated or not. Our assumption here is that grouping all of the sentences that contain one or more emotion annotations together can allow us to generate a model that can adequately separate sentences with one or more emotion annotations from those without.

Once a sentence has been identified as containing emotion annotations, the second stage of our system emits one or more target emotion labels. Due to limited time and resources for this effort, we decided to focus on the case of generating single emotion hypotheses, instead of multi-label methods. The majority of annotated sentences only have one emotion annotated, accounting for 86% of the total, as shown in Table 1. Given an initial scan of the training notes, we made the assumption that models developed for the single emotion case can be extended to multiple emotions.

In this case, we found that treating this as a multiclass classification problem outperformed using individual binary classifiers for each target emotion. This was likely due to the significant skew in the distribution over emotion annotations, with majority class labels such as *instructions* and *hopelessness* dominating the scoring function used to optimize the binary classifiers. Thus our second stage classifier was trained as a multiclass labeler.

In order to account for the remaining 14% of sentences that have multiple annotations, during training we treated these sentences as being individual instances of each emotion that was found. We experimented with using partially weighted instances, but found its performance to be poorer in comparison. In order to emit multiple emotion annotations at test time, we simply output the top scoring emotion and any the emotions whose scores were within 75% of the top emotion's score.

We initially experimented with using different sets of features for each of the two stages, but found that the features we experimented with lead to improvements in performance over the training set to be comparable for both stages. Thus we used the same set of features for both the first and second stages, with any differences noted below in the feature description.

Features

We now describe the types of features used by our system to characterize instances, along with the motivations for their inclusion. Our features were divided into three categories: lexical, psycholinguistic, and emotional sequence. All are applied on a per-sentence basis. The lexical features are derived from the orthographic representation of the sentences, while the psycholinguistic features map words and phrases encountered into psychologically valid dimensions. The emotional sequence features use both the ordering and placement of emotions to govern which emotions to emit at test time.

Lexically Derived Features

For given a sentence, we removed known English stop words, lowercased the sentence, and applied a whitespace tokenizer to segment the words, from which we extracted unigrams and bigrams, which were used directly as features for that instance. This is the "bag-of-words" approach commonly used for text classification, and we also use it as a baseline system for comparison with our other features. For the other non-baseline lexical features described here, we did not remove stop words, as they may be a significant component of a feature.

Previous studies have shown that part-of-speech (POS) can play a significant role in a variety of classification tasks involving populations with psychiatric or neurological issues, and can

discriminate between suicidal and non-suicidal language.[2,3] To obtain the POS tags for sentences, we used the Stanford Part-of-Speech tagger[4] to tag the tokens in the original sentence with their part-of-speech. For a given sentence, we collected the the frequencies of occurence of single POS tags and bigrams of tags, and incorporated these directly as features for our instances.

In order to capture the kind of actions described by nested expressions such as *"Please tell Jane to give my love to the children."*, we encode the lexeme and tense of the first and last verbs encountered in the sentence. This is used to encode both the syntactic and semantic heads found in a sentence. We also used the root verb from the typed dependency parse to augment this information, obtaining dependency parses from the Stanford parser.[5,6] The intent is for these features to highlight illocultionary acts that would characterize sentences labeled with *instructions*, allowing separation from language that would indicative of the more affective annotations.

During our analysis of the notes, we found that the way sentences began tended to govern which emotion annotations were assigned to that sentence. For example, those which contained the emotions *instructions* or *information* tended to be addresses to readers, such as *"To the police ..."* or *"To my wife ..."*. For example, sentences labeled with *hopelessness* tended to begin with the word *"I"*: out of 455 sentences marked with *hopelessness*, 156 of them started with *"I"* whereas only one began with *"Please"*. In comparison, of the 820 sentences labeled with *instructions*, 71 of them began with *"Please."* To capture these regularities, we specifically identify the first two words of each sentence as features.

Psycholinguistic Features

As the majority of the target annotations are expressions of emotions, we sought to incorporate information about the psychological and emotional content of the notes by using Linguistic Inquiry and Word Count[7] (LIWC). LIWC is a psycholinguistic resource that assigns one or more psychological categories such as *positive emotion* and *tentativeness* to individual words. For example, the word *"happy"* would be labeled with the categories *positive emotion* and *affect*. By scanning a text and assigning categories to applicable words in that text, one can derive an

aggregate signature of the psychological character of that text. LIWC currently has 80 categories.

In order to perform category assignment for a given sentence, LIWC perform a lexical match against its word to category dictionary. As such, LIWC is essentially performing a look-up, without conducting any part-of-speech identification nor sense disambiguation, and employs a few simple look-aheads to deal with a handful of ambiguous cases. In the authors' experience, when given a word, LIWC usually presumes that word's primary part-of-speech and word sense for assigning psychological and emotional dimensions. For example, the categories induced by the word *"cold"* are *percept* and *feel*, which would correspond to the adjective relating to the physical sensation of lowered temperature, instead of the adjective used to describe a person with little or no emotion, or the noun form used to describe an infection. Despite this apparent deficiency, a previous study found LIWC to better overall for identifying emotions, compared to similar psycholinguistic resources.[8]

For each sentence, we applied LIWC and used the returned counts directly as features. Because LIWC's analysis also included explicitly non-emotional categories that may be redundant with information already encoded by the POS tagger, such as the presence of pronouns and prepositions, we used only LIWC categories that contained emotional content.

One of the primary motivations for using psycholinguistic resources such as LIWC is to introduce additional knowledge that could help identify the rarer emotions. As shown in Table 2, the top eight emotions account for 90% of all annotations. This leaves the remaining seven emotions at risk of being overpowered, as the optimizer used to train the emotion classifier is likelier to favor the majority classes and neglect emitting the minority classes as hypotheses. We hoped to ameliorate this by introducing potentially strong signals into the featureset that correlate highly with just those minority classes. By doing this, the classifier's performance on those classes should be improved.

During develvopment, we found that LIWC tended to assign multiple labels to words that would ideally like to be identified using a single label. For example, words commonly associated with the *pride* emotion consistently mapped to the LIWC *affect*, *posemo*,

and *achieve* categories. This may introduce problems with the learner when dealing with another category that also scores high on a subset of those categories, such as *affect* and *posemo*. By using a single feature to tie those occurrences together, we hope to produce a stronger signal that the optimizer can use during classifier training. We introduced another feature which looked for specific combinations of LIWC categories over each word, and for matches found the corresponding single feature was added to that instance. We targeted the minority emotions *sorrow*, *pride*, and *happiness/peacefulness* with this feature.

Although LIWC profiles text along 80 dimensions, there are only three of these that we consider clearly relevant to the 8 "emotion" tags of this challenge. Those three categories are *affect*, *negemo* (negative emotion) and *posemo* (positive emotion), and these did not exhibit a very strong correspondance with the target emotions we wish to annotate with. We also found that more often than not, the targeted emotions were usually expressed by phrases instead of the presence of individual words.

To this end, we developed our own custom word and phrase lists that targeted the emotion annotations of interest. Like LIWC, these are applied over a source sentence, and enter as features the number of matches found in that sentence. These lists were developed using the training notes, as well based off of experience.

Emotional Sequence Features

During our data analysis, we observed that the sequence of emotion annotations tended to follow certain patterns. For example, we found that sentences annotated with *instructions* tended to precede those labeled with *information*, whereas sequences such as *thankfulness* followed by *anger* are comparatively much rarer. This notion of certain sequences "making sense" is similar to that of using sequence-based language models to identify coherent text, and the concept of discourse coherence, where a multi-sentence text is considered coherent if its arrangement of content allows it to convey its meaning. This also matches the intuition that authors of texts tend to exhibit regularity in transitions between the emotions they wish to convey. In order to capture this form of "emotional coherence," we employ a Markov model up to order two over the sequence of emotion annotations found in a note. As we observed from the training data, this

Table 3. Baseline system, performance by emotion and overall results.

Label	Precision	Recall	F1
NO EMOTION	0.6319	0.6463	0.6390
Abuse	0	0	0.000
Anger	0.0377	0.0180	0.0244
Blame	0.1406	0.1047	0.1200
Fear	0	0	0.000
Forgiveness	0	0	0.000
Guilt	0.4340	0.3977	0.4150
Happiness/peacefulness	0	0	0.000
Hopefulness	0.0392	0.0303	0.0342
Hopelessness	0.4732	0.5187	0.4949
Information	0.4412	0.3557	0.3939
instructions	0.6346	0.6855	0.6591
Love	0.6625	0.6262	0.6439
Pride	0	0	0.000
Sorrow	0	0	0.000
Thankfulness	0.4526	0.3924	0.4203

Notes: F1 = 0.3790; PRECISION = 0.3377; RECALL = 0.4318; N = 3225.

form of coherence tended to include the lack of annotations for a sentence, and we explicitly encode it as its own *no emotions* label.

For the first stage classifier, we group the presence of any emotions into a single *have emotions* label versus *no emotions*, as this improved first stage performance during development. For the second stage classifier we specifically identify the selected annotation. For sentences annotated with multiple emotions, we simply used the first emotion encountered in the annotation set. During training time, we used the gold emotion annotations to develop our sequence model, and during test time we evaluate using the emotions assigned by the classifier to previous sentences. We had attempted to train off of the annotations generated by our second stage classifier, but found the results to be poorer compared to using the gold annotations. For both stages, the frequency counts of the order one and two sequences are used as features.

In addition, we noticed that certain emotions tended to group in certain positions of the notes. To account for this behavior, we also included the current line number as a feature.

Evaluation and analysis

We now analyze features and performance based on our two stages: how well the system can identify whether emotions should be added or not, and how well it can guess the emotions. Assessment here was conducted using two-fold cross validation over the training notes.

We first note the performance of the baseline system, with a more in-depth view of scores by emotion, along with overall score over the training notes, given in Table 3. We include the lack of any emotion annotations as its own label, "NO EMOTION," in order to assess the performance of the first stage of our system. As noted before, the baseline system uses only the unigrams and bigrams found in each sentence as features.

Table 4. Full system, performance by emotion and overall results, lift/loss over baseline given in parentheses.

Label	Precision	Recall	F1
NO EMOTION	0.6255 (−0.0064)	0.7649 (+0.1186)	0.6882 (+0.0492)
Abuse	0 (0)	0 (0)	0 (0)
Anger	0.1212 (+0.0835)	0.0374 (+0.0194)	0.0571 (+0.0328)
Blame	0.3117 (+0.1711)	0.1491 (+0.0444)	0.2017 (+0.0817)
Fear	0.4 (+0.4)	0.0571 (+0.0571)	0.1 (0)
Forgiveness	0 (0)	0 (0)	0 (0)
Guilt	0.512 (+0.078)	0.3798 (-0.0179)	0.4361 (+0.0211)
Happiness/peacefulness	0 (0)	0 (0)	0 (0)
Hopefulness	0.2222 (+0.183)	0.0606 (+0.0303)	0.0952 (+0.0611)
Hopelessness	0.5592 (+0.086)	0.5584 (+0.0397)	0.5588 (+0.0639)
Information	0.5889 (+0.1477)	0.4569 (+0.1012)	0.5146 (+0.1207)
Instructions	0.6931 (+0.0585)	0.6865 (+0.0011)	0.6898 (+0.0308)
Love	0.7443 (+0.0818)	0.6586 (+0.0324)	0.6988 (+0.055)
Pride	0 (0)	0 (0)	0 (0)
Sorrow	0 (0)	0 (0)	0 (0)
Thankfulness	0.775 (+0.3224)	0.4079 (+0.0155)	0.5345 (+0.1141)

Notes: F1 = 0.4378; PRECISION = 0.4621; RECALL = 0.4159; N = 2270.

Table 5. Baseline with full set of lexical features, lift/loss over baseline given in parentheses.

Label	Precision	Recall	F1
NO EMOTION	0.6231 (−0.0088)	0.7511 (+0.1048)	0.6812 (+0.0421)
Abuse	0 (0)	0 (0)	0 (0)
Anger	0.125 (+0.0873)	0.018 (0)	0.0315 (+0.0071)
Blame	0.3596 (+0.2189)	0.1798 (+0.0751)	0.2397 (+0.1197)
Fear	0 (0)	0 (0)	0 (0)
Forgiveness	0 (0)	0 (0)	0 (0)
Guilt	0.5357 (+0.1018)	0.3715 (−0.0262)	0.4388 (+0.0237)
Happiness/peacefulness	0 (0)	0 (0)	0 (0)
Hopefulness	0.2857 (+0.2465)	0.0656 (+0.0353)	0.1067 (+0.0725)
Hopelessness	0.5401 (+0.067)	0.5197 (+0.0009)	0.5297 (+0.0348)
Information	0.485 (+0.0438)	0.3973 (+0.0416)	0.4368 (+0.0429)
Instructions	0.6583 (+0.0237)	0.7104 (+0.0249)	0.6834 (+0.0243)
Love	0.7386 (+0.0761)	0.6185 (−0.0077)	0.6732 (+0.0294)
Pride	0 (0)	0 (0)	0 (0)
Sorrow	0 (0)	0 (0)	0 (0)
Thankfulness	0.6744 (+0.2219)	0.3671 (−0.0253)	0.4754 (+0.0551)

Notes: F1 = 0.4217; PRECISION = 0.4341; RECALL = 0.4100; N = 2382.

For comparison, we show the performance of the full system in Table 4. We note any lift over baseline performance next to the entries for specific emotions, with gains given as positive values and losses as negative values. Overall, the full system achieves higher scores over the baseline, but is still unable to identify low frequency classes such as *forgiveness* or *sorrow*.

We give the results for systems using the baseline with the full set of lexical features in Table 5, the baseline with the psycholinguistic features activated in Table 6. Here, the baseline system we compare against is a "standard" text classification model that employs unigrams and bigrams, as described in the lexical features section. Any lift in performance over the baseline system is given in parentheses as a positive value, and any loss is given as a negative value.

We note that both the use of the full set of lexical features and the psycholinguistic features give gains

Table 6. Baseline system with psycholinguistic features, lift/loss over baseline given in parentheses.

Label	Precision	Recall	F1
NO EMOTION	0.6258 (−0.0061)	0.7313 (+0.0849)	0.6744 (+0.0354)
Abuse	0 (0)	0 (0)	0 (0)
Anger	0.16 (+0.1223)	0.0377 (+0.0197)	0.0611 (+0.0367)
Blame	0.2418 (+0.1011)	0.1257 (+0.0211)	0.1654 (+0.0454)
Fear	0.3333 (+0.3333)	0.1111 (+0.1111)	0.1667 (0)
Forgiveness	0 (0)	0 (0)	0 (0)
Guilt	0.4772 (+0.0432)	0.4237 (+0.026)	0.4488 (+0.0338)
Happiness/peacefulness	0 (0)	0 (0)	0 (0)
Hopefulness	0.2143 (+0.1751)	0.0882 (+0.0579)	0.125 (+0.0908)
Hopelessness	0.529 (+0.0558)	0.529 (+0.0103)	0.529 (+0.0341)
Information	0.5029 (+0.0617)	0.3772 (+0.0215)	0.4311 (+0.0372)
Instructions	0.6707 (+0.0361)	0.6794 (−0.0061)	0.675 (+0.016)
Love	0.6977 (+0.0351)	0.6613 (+0.0351)	0.679 (+0.0351)
Pride	0 (0)	0 (0)	0 (0)
Sorrow	0 (0)	0 (0)	0 (0)
Thankfulness	0.6588 (+0.2063)	0.3709 (−0.0215)	0.4746 (+0.0542)

Notes: F1 = 0.4140; PRECISION = 0.4060; RECALL = 0.4223; N = 2623.

Table 7. Baseline system with emotional sequence features, lift/loss over baseline given in parentheses.

Label	Precision	Recall	F1
NO EMOTION	0.628 (−0.0039)	0.7081 (+0.0618)	0.6657 (+0.0266)
Abuse	0 (0)	0 (0)	0 (0)
Anger	0.0526 (+0.0149)	0.0196 (+0.0016)	0.0286 (+0.0042)
Blame	0.2364 (+0.0957)	0.1512 (+0.0465)	0.1844 (+0.0644)
Fear	0.25 (+0.25)	0.0606 (+0.0606)	0.0976 (0)
Forgiveness	0 (0)	0 (0)	0 (0)
Guilt	0.4465 (+0.0125)	0.4492 (+0.0515)	0.4479 (+0.0328)
Happiness/peacefulness	0 (0)	0 (0)	0 (0)
Hopefulness	0.3333 (+0.2941)	0.087 (+0.0567)	0.1379 (+0.1037)
Hopelessness	0.5179 (+0.0447)	0.5066 (−0.0122)	0.5121 (+0.0172)
Information	0.4836 (+0.0424)	0.4488 (+0.0931)	0.4655 (+0.0717)
Instructions	0.6761 (+0.0415)	0.6877 (+0.0022)	0.6818 (+0.0228)
Love	0.7056 (+0.0431)	0.6559 (+0.0297)	0.6799 (+0.036)
Pride	0 (0)	0 (0)	0 (0)
Sorrow	0 (0)	0 (0)	0 (0)
Thankfulness	0.6598 (+0.2072)	0.4129 (+0.0205)	0.5079 (+0.0876)

Notes: F1 = 0.4184; PRECISION = 0.4438; RECALL = 0.3957; N = 2249.

over the baseline system. However these gains were primarily over the the top eight most frequent emotions. For the seven least frequent emotions, these were still essentially neglected by the system, and the psycholinguistic features only managed to identify a handful of *fear* annotations. Closer examination of the confusion matrix shows that these are often misguessed as other emotions: sentences labeled with *forgiveness* are mis-labeled with *guilt*, and those labeled with *sorrow* are usually mislabeled with *hopelessness*.

The performance of the baseline system with the emotion coherence features is given in Table 7. What is interesting to note is even with a simple sequence model of emotional coherence, we see an overall gain in performance.

Examination of the confusions derived from testing on the same data as was trained on showed exhibited strong overfitting over all systems.

Table 8. System performance over test set.

Model	Precision	Recall	F1	Guesses
Baseline	0.4623	0.4631	0.4627	1274
Full system	0.5000	0.4686	0.4838	1192
Full system + Bootstrap	0.5114	0.4764	0.4933	1185

Performance of the baseline and full system is given in Table 8.[1]

Missing annotations and bootstrapping

One of our observations during an analysis of the notes was that annotator fatigue appeared to play an issue. There were numerous cases where a sentence contained no emotion annotation, yet given the language observed we expected an annotation to be present.

Given that assumption that our training data is partially labeled, we tested our system by treating the labeled data as seeds for bootstrapping. We performed just a single iteration of bootstrapping over the training set, this being equivalent to an early stop employed by "cautious" learning approaches over unsupervised data.[9] We found that this gave a small increase in precision and recall, amounting to nearly an additional point in F1 over the test set (Table 8).

Conclusion and Future Work

We have described how a variety of lexical, psycholinguistic, and emotional sequence features can augment a baseline text-classification system. A clear

---

[1]We identified a bug in our system after submisssion, and performance measures here reflect those of the fixed system.

area for future improvements is in the handling of the multi-label cases, such as employing a classifier that specifically targets multiple labels. And given the signal derived from the emotional sequence model, improvements to both how we model multiple emotions and the emotional transitions are warranted.

Another area of improvement would be to address the seven minority emotion annotations, which constituted only 10% of all the annotations introduced. We could certainly employ methods that can incorporate a small number of positive instances, such as using an instance-based classifier. However, the lack of available training data for these classes would argue for more training data to cover those cases, or some deterministic rules in our system to account for them.

We also found that misspellings and poor grammar tended to be present for in significant number of notes. In addition to introducing extra sparsity into the feature space, this would also present a problem for resources, such as LIWC, that rely on lexical matches. Also, even if a more reliable sentence segmentation algorithm were employed for future emotion annotations, the presence of grammatical errors could still impact what is deemed a sentence.

One point of interest here is the under performance of the psycholinguistic resources. Analysis of the notes showed that in most of the cases we observed, the emotions of interest were commonly associated with phrases instead of single words. This is particularly true of the *abuse* emotion, which is more of affect-laden judgement about another's behavior and consequently may be harder to characterize using a bag of words model. This also applied to the case of sentences containing multiple annotations, where we observed that constituent phrases tended to account for only a single emotion. For example, the sentence *"Please forgive me, but I can't go on like this."* would be annotated with the emotions *guilt* and *hopelessness*. However, based off observations from the single emotion sentences, we would argue that the phrase *"Please forgive me"* is directly responsible for the *guilt* emotion, and the clause *"I can't go on like this"* encompasses *hopelessness*.

Certainly increasing the range of phrasings covered by our phraselists would be one way to account for this. However, given the apparent phrase-based nature of the problem, we would argue that a system that worked on the token to phrase level would be more appropriate for this task than one that works on a sentence level. Indeed, during development it became apparent that viewing the fundamental task as that of information extraction, instead of text classification, may have been a better fit for this task. In general, information extraction approaches treat the hypotheses of interest as applying over word spans, instead of entire sentences. Given this, we would recommend that future annotations of this form be performed at the word span level.

Acknowledgements

Disclosures

References

1. Pestian JP, Matykiewicz P, Linn-Gust M, Wiebe J, Cohen K, Brew C, et al. Sentiment analysis of suicide notes: A shared task. 2012;5 (Suppl. 1):3–16.
2. Matykiewicz P, Duch W, Pestian J. Clustering semantic spaces of suicide notes and newsgroups articles. In *Proceedings of the BioNLP* 2009 *Workshop*, pages 179–184, Boulder, Colorado, June 2009. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W09–1323.
3. Peintner B, Jarrold W, Vergyri D, Richey C, Tempini ML, Ogar J. Learning diagnostic models using speech and language measures. In: *Conference Proceedings IEEE Eng Med Bio Society*, 2008.
4. Toutanova K, Klein D, Manning CD, Singer Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In: *Proceedings of HLT-NAACL*, 2003:252–59.
5. de Marneffe MC, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In *LREC* 2006, 2006.

6. Klein D, Manning CD. Accurate unlexicalized parsing. In: *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003.

7. Pennebaker JW, Booth RJ, Francis ME. Linguistic inquiry and word count (http://www.liwc.net), 2007.

8. Bantum E, Owen JE. Evaluating the validity of computerized content analysis programs for identification of emotional expression in cancer narratives. *Psychological Assessment*. Mar 2009;21:79–88.

9. Collins M, Singer Y. Unsupervised models for named entity classification. In: *In Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*, 1999:100–10.