AFFECTIVE LANGUAGE MODEL ADAPTATION VIA CORPUS SELECTION

Nikolaos Malandrakis[1], *Alexandros Potamianos*[1], *Kean J. Hsu*[2], *Kalina N. Babeva*[2],
Michelle C. Feng[2], *Gerald C. Davison*[2], *Shrikanth Narayanan*[1]

[1]Signal Analysis and Interpretation Laboratory (SAIL), USC, Los Angeles, CA 90089, USA
[2]Laboratory for Cognitive Studies in Clinical Psychology, USC, Los Angeles, CA 90089, USA
malandra@usc.edu, potam@telecom.tuc.gr, keanhsu@usc.edu, babeva@usc.edu,
michelcf@usc.edu, gdaviso@usc.edu, shri@sipi.usc.edu

ABSTRACT

Motivated by methods used in language modeling and grammar induction, we propose the use of pragmatic constraints and perplexity as criteria to filter the unlabeled data used to generate the semantic similarity model. We investigate unsupervised adaptation algorithms of the semantic-affective models proposed in [1, 2]. Affective ratings at the utterance level are generated based on an emotional lexicon, which in turn is created using a semantic (similarity) model estimated over raw, unlabeled text. The proposed adaptation method creates task-dependent semantic similarity models and task-dependent word/term affective ratings. The proposed adaptation algorithms are tested on anger/distress detection of transcribed speech data and sentiment analysis in tweets showing significant relative classification error reduction of up to 10%.

***Index Terms*—** emotion, affect, affective lexicon, polarity detection, language understanding.

1. INTRODUCTION

The analysis of language affective content is a significant part of many applications involving written or verbal speech, such as sentiment analysis [3], news headlines analysis [4] and emotion recognition from multimedia streams (audio, video, text) [5, 6]. Affective text analysis can happen at various levels, targeting different lexical units: words, phrases, sentences, utterances, as appropriate to the task. Analyzing the content of utterances typically involves compositional [7] models of affect, that express the meaning of the utterance through some combination of the meanings (typically affective ratings) of the words they contain. Word ratings are provided by affective lexica, either manually annotated, such as Affective norms for English Words (ANEW) [8] or, more typically, automatically expanded lexica such as SentiWordNet [9] and WORDNET AFFECT [10]. These word ratings are then combined through a variety of methods, making use of part-of-speech tags [11], sentence structure [12] or hand-tuned rules [13].

A common problem for affective language analysis is the large variety of topics and discourse patterns that may be observed and their effect on content interpretation. Different domains can contain text of different topics, leading to words being used with different senses, or text created using different styles of speech/writing, e.g. informal or self-referential. This poses challenges since most popular resources are domain-agnostic and therefore sub-optimal if the task is focused on a narrow domain. There are two main solutions to this problem: 1) topic modeling of general purpose data and 2) domain adaptation via data selection. *Topic modeling* [14, 15] typically aims to represent the meaning of words (and by extension sentences and beyond) through a probabilistic mixture of topic-specific models, in which case the affective content is estimated over all topics. While these models show promise, they do not fit particularly well with other computational frameworks: in the recent SemEval sentiment analysis challenge [3] virtually no submissions used topic modeling, opting for methods based on affective lexica. Here instead we take the direct domain adaptation approach that has been very successful in the language modeling and grammar induction literature [16, 17].

Our proposed method involves automatically adapting an affective lexicon in order to better suit a task. Virtually all automatically generated lexica are created based on some form of word similarity and the assumption that semantic similarity implies affective similarity. Therefore if we can estimate domain-dependent word similarity scores then we can create domain-dependent affective word/term ratings. Our method of lexicon expansion [18], unlike popular alternatives [9, 10], is purely data-driven, utilizing web-harvested data and estimating similarity scores through statistics. A simple, yet general and efficient way to adapt to a specific domain is to filter the data used to estimate the semantic similarity model. The data selection process we propose is inspired by similar methods of harvesting data from the web used for language modeling [16] and grammar induction [17]. Given a small amount of in-domain data we can, in an unsupervised fashion, select similar data from a large corpus through the use of pragmatic constraints introduced in [17] and perplexity, leading to a smaller corpus that is more relevant to the task. Using this corpus we can create domain-specific similarities and affective ratings. Compared to previous research and topic modeling our approach differs in that it generates a single model rather than a mixture of models. It also results in an affective lexicon, a resource that is more versatile, since it can fit within most computational frameworks. Next we outline the basic semantic-affective model of [1, 2], detail how to expand it to any type of target labels (e.g., distress, anger, sentiment), as well as to adapt it using: 1) adaptation of the semantic model through utterance selection and 2) direct adaptation of the semantic-affective map. The proposed methods are evaluated on affective tasks of both speech transcribed data and text twitter data.

2. CREATING AFFECTIVE RATINGS

To generate affective ratings for utterances we use a compositional framework. The utterance is broken into a bag of all words and bigrams it contains, affective ratings for them are estimated from a lexicon and finally statistics of these word/bigram ratings are combined into a sentence rating.

The bootstrap lexicon we use is automatically expanded using a method first presented in [18] and expanded in [1]. It builds on [2]. We start from a manually annotated lexicon, preferably annotated in continuous affective scales and pick, from the the lexicon, the words corresponding to the most extreme ratings, e.g., for valence we pick the most positive and most negative words in the lexicon, and use them as dimensions to define a semantic model, a space representing semantic similarities to these seed words. Then we use a large corpus to calculate statistics and estimate semantic similarity metrics that will allow us to place any words or bigrams in the semantic space. We also define an affective model, a space, in this case one of arousal–valence–dominance, where we aim to place any new word. The mapping from the semantic model to the affective one is trained on the annotated lexicon using Least Squares Estimation (LSE) and is a simple linear function

$$\hat{v}(w_j) = a_0 + \sum_{i=1}^{N} a_i \; v(w_i) \; d(w_i, w_j), \tag{1}$$

where w_j is the word whose affect we aim to characterize, $w_1 ... w_N$ are the N seed words, $v(w_i)$ is the affective rating for seed word w_i, a_i is the (trainable) weight corresponding to word w_i and $d(w_i, w_j)$ is a measure of semantic similarity between words w_i and w_j. While $d(w_i, w_j)$ may be any estimate of semantic similarity, we have found that the cosine similarity between the binary weighted context vectors of w_i and w_j performs best [1]. An overview of the lexicon expansion model can be seen in Fig. 1. For more details see [1].

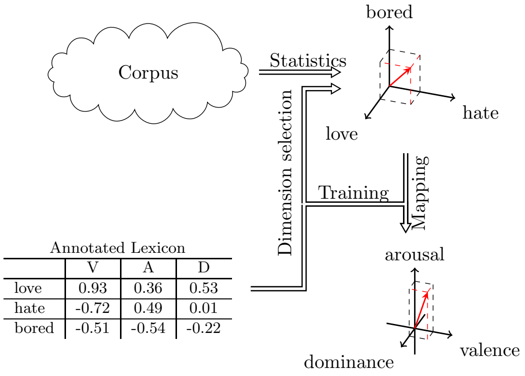


Fig. 1. Overview of the lexicon expansion method

The final step is the mapping of the semantic-affective model in (1) to various categorical labels at the sentence or paragraph level, e.g., sentiment, distress, anger, politeness, frustration. Typically this last step is achieved via a classifier that takes as input the 3-D affective rating of each token (unigrams and bigrams) and produces sentence/utterance level statistics of these ratings (e.g., mean, median, variance). The feature fusion scheme is trained for each specific categorical label separately[1].

3. ADAPTATION

Generating the values of the similarity metric used in (1) requires a text corpus of sufficient size so as to contain multiple instances of the words and bigrams we want to generate ratings for. Size requirements like this one have driven researchers to the web, which should contain sufficient data for virtually any need. However it is still necessary to sample the web, in order to collect appropriate data, e.g., we may harvest only data from twitter for some tasks.

Instead of adapting the semantic-affective space directly[2] we choose to adapt the semantic similarity scores by reestimating semantic similarity on a subset of the original corpus that better matches the content and style of our in-domain data. Thus adaptation boils down to an utterance selection method. In this work, motivated by work in language modeling and grammar induction, we utilize two criteria to perform utterance selection: pragmatic constraints and perplexity.

Pragmatic constraints are terms or keywords that are highly characteristic of a domain. For example when generating a domain independent corpus we would search for “delay”, however if we know that the application is related to air travel then we can use terms that are highly characteristic of the domain, like “flight” and “plane”. By constraining our sub-corpus to contain one of these words we will get a sub-corpus that is conditioned on the domain and in turn allow us to estimate domain dependent probabilities and other metrics, e.g., semantic similarity. While in this example the pragmatic words are content words, that is not necessarily the case. The target application may be better characterized by stylistic elements, e.g., interview transcripts will contain many self-references which may lead to the word “I” being highly characteristic. Identifying these characteristic terms can be done via comparing an in-domain corpus with a generic corpus. Intuitively, highly characteristic terms should appear relatively more often in the in-domain corpus and also appear on multiple separate samples (utterances), or equivalently should have a high value as proposed in [17]:

$$D(w) \frac{P_{in}(w)}{P_{out}(w)}, \tag{2}$$

where $D(w)$ the number of in-domain samples (sentences, documents or otherwise) the term occurs in, $P_{in}(w)$ the probability of the term in the in-domain corpus and $P_{out}(w)$ the probability of the term in the generic corpus.

Perplexity is a popular method of estimating the degree of fit between two corpora, by generating a language model on one and calculating the perplexity on the other. In this context, we can generate a language model using the in-domain corpus and use it to evaluate each instance contained in the generic corpus [16]. Instances that are lexically more similar to the instances in the in-domain corpus will be assigned lower perplexity scores. Therefore we can apply a threshold on perplexity to detect if a new instance should be included to our task-dependent corpus. Once the corpus is selected using pragmatic constraints and/or perplexity thresholding the semantic similarity metrics are re-estimated on the selected sub-corpus.

Instead of or in addition to adapting the semantic similarity model $d(.,.)$ in (1) one could adapt directly the semantic-affective mapping, i.e., the parameters a_i in (1) using in domain data (or more

[1] Provided there is enough in domain data one could also build a direct mapping from the semantic similarity space to the label space, i.e., not use the affective rating as an intermediate mapping step. Alternatively on could combine the semantic-affective-label model with a semantic-label model. Given the limited space we do not present results on such model combinations here.

[2] Such adaptation is possible however constraints have to be posed on the semantic space to be able to perform it. Unfortunately, the semantic space (although often represented as an inner product space in distributed semantic models (DSMs)) is far from metric, i.e., the triangular inequality is often violated. Adapting the corpus used to estimate semantic similarity is an elegant way to bypass the problem of adapting the non-metric semantic space.

realistically mixing in-domain and general purpose data) as outlined in [18]. This method is also evaluated and compared to semantic space adaptation for the twitter data.

4. EXPERIMENTAL PROCEDURE

The main word corpus we use to train the lexicon creation algorithm is the *Affective Norms for English Words* (ANEW) dataset. ANEW consists of 1034 words, rated in 3 continuous dimensions of arousal, valence and dominance.

To train sentence-level models and evaluate their performance we use subsets of the Articulated Thoughts in Simulated Situations [19] (ATSS) paradigm corpus and the SemEval2013 Task 2 [3] twitter dataset. The ATSS paradigm corpus is composed of manually transcribed sessions of a psychological experiment, where participants are presented with short segments of emotion-inducing scenarios and respond, typically with a few sentences per utterance. These utterances are manually annotated on multiple scales. For these experiments we use a subset of 1176 utterances and binary labels of anger (522 positive) and distress (445 positive). The twitter corpus contains individual tweets annotated as positive, negative and neutral. For these experiments we use the training set, composed of 9816 tweets and containing 1493 negative, 4649 neutral and 3674 positive samples.

In order to evaluate various methods of utterance selection we need a starting domain-independent corpus. To create one we use the vocabulary of English packaged in the aspell spellchecker for English, containing 135,433 words, pose a query for each of them to the Yahoo! search engine and collect the snippets (short representative excerpts of the document shown under each result) of the top 500 results. Each snippet is usually composed of two sentences: title and content. The corpus contains approximately 117 million sentences.

This corpus, as well as filtered versions thereof, is used to calculate statistics and generate the required semantic similarity metrics to be used by the word/term model. The model itself is created by selecting seed words from ANEW and training on the entire ANEW corpus and then used to generate arousal, valence and dominance ratings for all words and bigrams contained in the evaluated utterance corpora.

Utterance level features are created by calculating word and bigram rating statistics across each utterance. The statistics used are: cardinality, minimum, maximum, range (maximum minus minimum), extremum (value furthest from zero), sum, average and standard deviation. Statistics are calculated across all terms and subgroups based on rough part-of-speech tags: verbs, adjectives, adverbs, nouns and their combinations of two three and four. For example, one feature may be the maximum arousal across all verbs and adjectives.

It should be noted that up to and including feature extraction we are using the affective dimensions of valence, arousal and dominance. While these dimensions are not the same as the utterance-level labels, they should be capable of representing them, e.g., anger should fall within the negative valence, high arousal, high dominance part of the space. The task of moving these ratings to the desired affective representation is handled by the supervised sentence model. The model, that uses the extracted features after selection, is a Naive Bayes Tree, a decision tree with a Naive Bayes classifier on each node.

5. RESULTS

Next we present the baseline results (general purpose corpus used for semantic similarity estimation), as well as the adaptation results using pragmatic constants and/or perplexity thresholding. To select the words that will form the pragmatic constraints we use (2) with the in-domain corpus being the evaluated utterance corpus and the generic corpus being the web-harvested 117m sentences. Using these we can score and rank every word in the in-domain corpus, however we do not know how many of these words we should pick. Data selection will result in a corpus that may be more salient to the task, but will also be smaller[3]. In order to keep a fairly large corpus, we keep the top-20 words for each training corpus and use them to filter the original 117m sentences.

To filter by perplexity we train trigram language models (Witten-Bell smoothing) on the in-domain corpora and use them to calculate perplexity for each of the sentences contained in the generic corpus. As previously there is no optimal value of perplexity we can aim for, since a lower threshold will lead to a smaller corpus. For these experiments we use thresholds of 100, 300, 1000 and 6000. Perplexity thresholding is applied to the original 117m sentences or to a corpus already filtered via pragmatic constraints. The sizes of the filtered corpora generated are shown in Table 1. All filtered corpora are substantially smaller than the initial corpus and as the perplexity constraint gets stricter corpora can become very small. However, even a corpus of fifty thousand sentences is very large compared to most annotated in-domain corpora available.

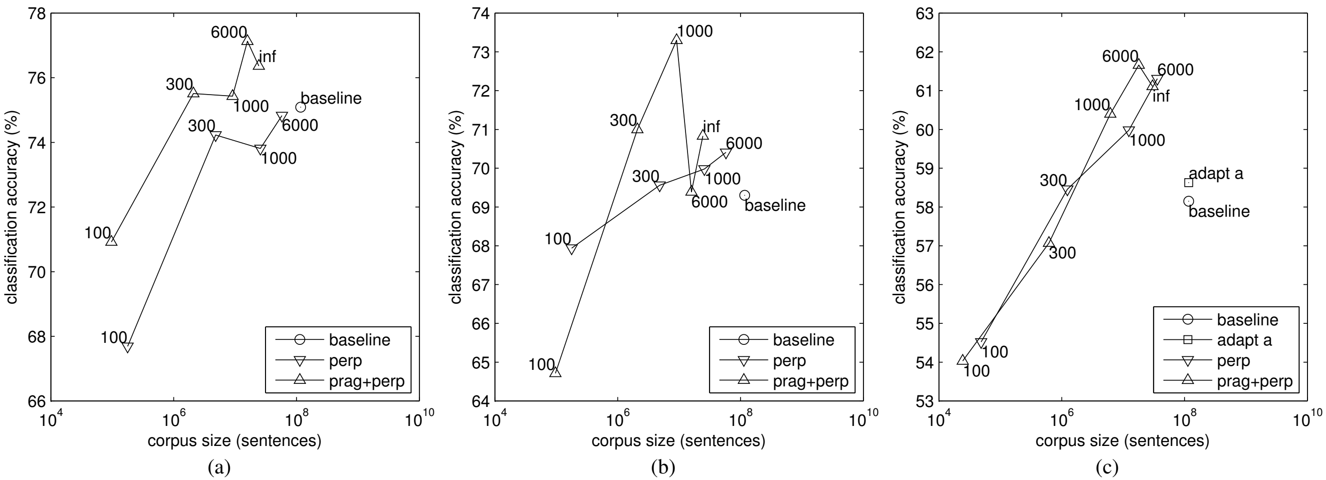
Table 1. Corpus size after pragmatic constraints and/or perplexity thresholding has been applied, for the ATSS and Twitter experiments.

Pragmatic Constraints	Perplexity Threshold	Sentences	
		ATSS	Twitter
no	-	116,749,758	116,749,758
no	100	177,786	48,868
no	300	4,837,935	1,241,524
no	1000	25,786,774	12,412,022
no	6000	57,932,887	36,044,486
yes	-	24,432,892	30,193,306
yes	100	96,768	24,434
yes	300	2,096,241	620,762
yes	1000	9,116,490	6,206,011
yes	6000	15,907,177	18,022,243

These corpora are used to generate semantic similarity estimates and create emotional ratings for all bigrams and unigrams in all utterances. The term model uses the 600 most extreme words in ANEW as seed words [1]. It should be noted that while the similarities are re-evaluated for each different corpus, the semantic-affective map is not: it is trained on baseline corpus of 117 million sentences and used as-is in all other cases.

To set baselines we use a domain-independent lexicon model trained using all 117m sentences. The model is, as shown in [1], very accurate reaching a Pearson correlation to the ground truth of 0.87, so it is a good representation of what a domain-independent model can do. In the case of twitter, we also compare against the supervised adaptation of the semantic-affective model proposed in [18]. This method performs no corpus selection, but rather re-trains

[3] We have shown [1] that performance increases with corpus size.



(a) (b) (c)

Fig. 2. Classification accuracy as a function of the size of the corpus used for lexicon creation, using perplexity, pragmatic constraints and perplexity or neither to generate the corpus. The data point labels correspond to perplexity threshold values. Performance shown for (a) ATSS anger, (b) ATSS distress and (c) TWITTER sentiment.

the a_i coefficients in (1). Words in the training utterances are assigned ratings equal to the average of all utterances they occur in and then are used as training samples to re-train the lexicon model. For twitter we only re-train the valence model since sentiment polarity is very similar to valence and set the negative/neutral/positive values as $-1, 0, 1$ valence.

To evaluate on the ATSS paradigm corpus we attempt prediction of the binary values of anger and distress using 10-fold cross-validation. The performance achieved, in terms of classification accuracy as a function of corpus size, is shown in Fig. 2. The baseline performance, achieved without any use of filtering, is 75% for anger and 69% for distress, with chance baselines of 56% and 62% respectively. Using perplexity alone can lead to an improvement over the baseline, however that improvement is much less compared to that achieved by using pragmatic constraints, with or without (infinite) a perplexity threshold. The combination of pragmatic constraints and perplexity results in a notable improvement of baseline, reaching accuracies of 77% and 73% respectively. Of note is the difference in optimal perplexity thresholds between the dimensions of anger and distress, indicating that the sample labels could (or perhaps should) be used as part of the filtering process.

To evaluate on Twitter we attempt prediction of the ternary value of sentiment using 10-fold cross-validation, the performance of which is shown in Fig. 2. The baseline performance is 58%, whereas the chance baseline is 47%. As is the case with ATSS anger, we see a substantial improvement when using a combination of pragmatic constraints and perplexity, reaching a peak of 62% or 4% over baseline. In this case the improvement gained by including pragmatic constraints rather than using just perplexity is much less pronounced than in the case of ATSS, however that is probably the result of picking a sub-optimal number of pragmatic words as constraints. The supervised adaptation of the a_i coefficients improves on the baseline, however the difference is very small, particularly when compared to the results of the unsupervised method.

Performance is improved notably in all cases showing the validity of the main idea of adapting the semantic similarity model. Pragmatic constraints seem a better selection criterion than perplexity, though peak performance is achieved by combining the two.

Finally, we investigate the use of a mixture of the adapted models used above and the task-independent model. To do that we take a weighted linear combination of the models $(w \cdot d_{in}(.,.) + (1 - w) \cdot d_{out}(.,.))$, where $d_{in}(.,.)$ the semantic similarity estimated over the filtered corpus and $d_{in}(.,.)$ the semantic similarity estimated over the task-independent 117m sentence corpus. The maximum performance achieved, as well as the corresponding weights w are shown in Table 2. As expected, the better performing adapted models get weighted more in the mixture (typically 80-20%). Combing the in-domain and out-of-domain models provides very little benefit in terms of maximum performance, however it increases robustness considerably, smoothing out the performance shown in Fig. 2. All but the worst performing in-domain models can achieve similar performance levels when in a mixture, though only the better in-domain models are assigned high weights.

Table 2. Performance for each experiment using linear combinations of the generic and adapted lexicon models. Presented is the maximum accuracy achieved in each case, as well as the parameters of the adapted model and the weight w assigned to it.

Experiment	Pragmatic Constraints	Perplexity Threshold	w	acc.
ATSS Anger	yes	6000	0.8	77.7%
ATSS Distress	yes	1000	0.8	73.9%
Twitter Sentiment	yes	1000	0.9	62.1%

6. CONCLUSIONS

We proposed a method of adapting an affective lexicon generation method to specific tasks through the use of corpus selection, as part of a system that generates utterance-level affective ratings. The method was shown to provide notable improvements in prediction accuracy on speech and twitter datasets. Future work should focus on finding optimal filtering parameters, number of pragmatic words and perplexity thresholds, as well as the role of labels in the corpus selection process. We will also investigate how to optimally combine various adaptation methods at the semantic similarity level, semantic-affective map, affective-label map, as well as the sampling of the semantic-affective space via seed work selection.

Acknowledgements: Prof. Alexandros Potamianos was partially supported by an Aristeia GSRT grant. Portions of this work have been performed under the EU-IST FP7 project PortDial.

7. REFERENCES

[1] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, “Distributional semantic models for affective text analysis,” *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 21, no. 11, pp. 2379–2392, 2013.

[2] P. Turney and M. L. Littman, “Unsupervised Learning of Semantic Orientation from a Hundred-Billion-Word Corpus. Technical report ERC-1094 (NRC 44929),” National Research Council of Canada, 2002.

[3] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson, “Semeval-2013 task 2: Sentiment analysis in twitter,” in *Proceedings of *SEM*, 2013, pp. 312–320.

[4] C. Strapparava and R. Mihalcea, “Semeval-2007 task 14: Affective text,” in *Proc. SemEval*, 2007, pp. 70–74.

[5] C. M. Lee and S. Narayanan, “Toward detecting emotions in spoken dialogs,” *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 2, pp. 293–303, 2005.

[6] C. M. Lee, S. Narayanan, and R. Pieraccini, “Combining acoustic and language information for emotion recognition,” in *Proc. ICSLP*, 2002, pp. 873–876.

[7] F. J. Pelletier, “The principle of semantic compositionality,” *Topoi*, vol. 13, pp. 11–24, 1994.

[8] M. Bradley and P. Lang, “Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings. Technical report C-1,” The Center for Research in Psychophysiology, University of Florida, 1999.

[9] A. Esuli and F. Sebastiani, “Sentiwordnet: A publicly available lexical resource for opinion mining,” in *Proc. LREC*, 2006, pp. 417–422.

[10] C. Strapparava and A. Valitutti, “WordNet-Affect: an affective extension of WordNet,” in *Proc. LREC*, 2004, vol. 4, pp. 1083–1086.

[11] F.-R. Chaumartin, “UPAR7: A knowledge-based system for headline sentiment tagging,” in *Proc. SemEval*, 2007, pp. 422–425.

[12] A. Andreevskaia and S. Bergler, “CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging,” in *Proc. SemEval*, 2007, pp. 117–120.

[13] K. Moilanen, S. Pulman, and Y. Zhang, “Packed feelings and ordered sentiments: Sentiment parsing with quasi-compositional polarity sequencing and compression,” in *Proc. WASSA Workshop at ECAI*, 2010, pp. 36–43.

[14] C. Lin and Y. He, “Joint sentiment/topic model for sentiment analysis,” in *Proceedings of the 18th ACM conference on Information and knowledge management*, 2009, CIKM, pp. 375–384.

[15] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang Zhai, “Topic sentiment mixture: modeling facets and opinions in weblogs,” in *Proceedings of the 16th international conference on World Wide Web*, 2007, WWW ’07, pp. 171–180.

[16] T. Ng, M. Ostendorf, Mei-Yuh Hwang, Manhung Siu, Ivan Bulyko, and Xin Lei, “Web-data augmented language models for mandarin conversational speech recognition,” in *Proceedings of ICASSP*, 2005, vol. 1, pp. 589–592.

[17] Ioannis Klasinas, Alexandros Potamianos, Elias Iosif, Spiros Georgiladakis, and Gianluca Mameli, “Web data harvesting for speech understanding grammar induction,” in *Proceedings of Interspeech*, 2013, pp. 2733–2737.

[18] N. Malandrakis, A. Potamianos, E. Iosif, and S. Narayanan, “Kernel models for affective lexicon creation,” in *Proc. Interspeech*, 2011, pp. 2977–2980.

[19] K.J. Hsu, K.N. Babeva, M.C. Feng., J.F. Hummer, and G.C. Davison, “Experimentally induced distraction impacts cognitive but not emotional processes in cognitive assessment,” *in Submission*.