HLT-NAACL 2006

Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics

Proceedings of the Doctoral Consortium

Matt Huenerfauth and Bo Pang
Doctoral Consortium Chairs
Mitch Marcus, Faculty Advisor

June 4-9, 2006
New York City, USA

Published by the Association for Computational Linguistics
http://www.aclweb.org

Table of Contents

Preface

We are pleased to present the 10 papers that have been accepted for presentation at the first HLT/NAACL Doctoral Consortium (New York City, June 2006). The goal of this event is to create an opportunity for a group of senior Ph.D. students to discuss and explore their research and career objectives with a panel of researchers in the fields of natural language processing, speech technology, and information retrieval. The event is also an opportunity for students to gain exposure for their work among the entire HLT/NAACL research community.

On June 4, 2006, the 10 participating students share their research in the form of short "job talk" presentations and receive feedback from a panel of researchers drawn from industry and academia. The day also includes a panel presentation by established researchers on the topic of "How to best present yourself on the academic/industrial job market." We would like to thank all of the researchers who have volunteered to serve on panels throughout the day.

Students also participate in a poster session held during the main HLT/NAACL conference and have their professional biography, research abstract, and photograph included in a face book to be distributed to all attendees of HLT/NAACL-2006.

This year, we received 18 submissions representing 8 countries. We would like to thank all of the authors who have submitted their work to this event.

We are grateful for the help and support of our faculty advisor Mitch Marcus and the HLT/NAACL-2006 main conference organizers, especially Bob Moore, Satoshi Sekine, Brian Roark, Jennifer Chu-Carroll, Mark Sanderson, Jeff Bilmes, Ed Hovy, Patrick Pantel, and Priscilla Rasmussen. We are also grateful for the advice we've received from Lillian Lee during the planning of this event. Finally, we gratefully acknowledge the National Science Foundation and Microsoft Corp., whose support will make it possible to provide financial assistance to the presenters.

Matt Huenerfauth and Bo Pang

ORGANIZERS:

Co-Chairs:

Matt Huenerfauth, University of Pennsylvania
Bo Pang, Cornell University

Faculty Advisor:

Mitch Marcus, University of Pennsylvania

DOCTORAL CONSORTIUM WEBSITE:

http://www.cis.upenn.edu/proj/hlt-naacl-2006-dc/

Incorporating Gesture and Gaze into Multimodal Models of Human-to-Human Communication

Lei Chen
Dept. of Electrical and Computer Engineering
Purdue University
West Lafayette, IN 47907
chenl@ecn.purdue.edu

Abstract

Structural information in language is important for obtaining a better understanding of a human communication (e.g., sentence segmentation, speaker turns, and topic segmentation). Human communication involves a variety of multimodal behaviors that signal both propositional content and structure, e.g., gesture, gaze, and body posture. These non-verbal signals have tight temporal and semantic links to spoken content. In my thesis, I am working on incorporating non-verbal cues into a multimodal model to better predict the structural events to further improve the understanding of human communication. Some research results are summarized in this document and my future research plan is described.

1 Introduction

In human communication, ideas tend to unfold in a structured way. For example, for an individual speaker, he/she organizes his/her utterances into *sentences*. When a speaker makes errors in the dynamic speech production process, he/she may correct these errors using a *speech repair* scheme. A group of speakers in a meeting organize their utterances by following a *floor control* scheme. All these structures are helpful for building better models of human communication but are not explicit in the spontaneous speech or the corresponding transcription word string. In order to utilize these structures, it is necessary to first detect them, and to do so as efficiently as possible. Utilization of various kinds of knowledge is important; For example, lexical and prosodic knowledge (Liu, 2004; Liu et al., 2005) have been used to detect structural events.

Human communication tends to utilize not only speech but also visual cues such as gesture, gaze, and so on. Some studies (McNeill, 1992; Cassell and Stone, 1999) suggest that gesture and speech stem from a single underlying mental process, and they are related both temporally and semantically. Gestures play an important role in human communication but use quite different expressive mechanisms than spoken language. Gaze has been found to be widely used in coordinating multi-party conversations (Argyle and Cook, 1976; Novick, 2005). Given the close relationship between non-verbal cues and speech and the special expressive capacity of non-verbal cues, we believe that these cues are likely to provide additional important information that can be exploited when modeling structural events. Hence, in my Ph.D thesis, I have been investigating the combination of lexical, prosodic, and non-verbal cues for detection of the following structural events: *sentence units*, *speech repairs*, and *meeting floor control*.

This paper is organized as follows: Section 1 has described the research goals of my thesis. Section 2 summarizes the efforts made related to these goals. Section 3 lays out the research work needed to complete my thesis.

2 Completed Works

Our previous research efforts related to multimodal analysis of human communication can be roughly grouped to three fields: (1) multimodal corpus col-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 211–214,
New York, June 2006.

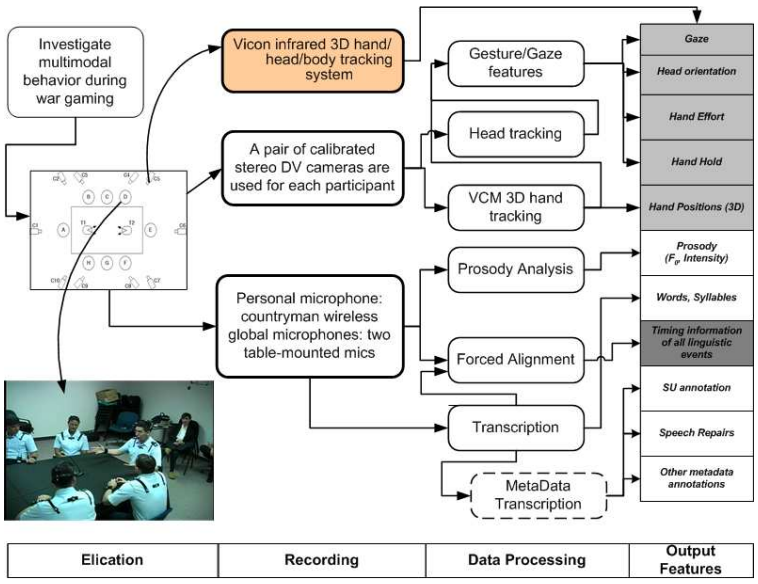


Figure 1: VACE meeting corpus production

lection, annotation, and data processing, (2) measurement studies to enrich knowledge of non-verbal cues to structural events, and (3) model construction using a data-driven approach. Utilizing non-verbal cues in human communication processing is quite new and there is no standard data or off-the-shelf evaluation method. Hence, the first part of my research has focused on corpus building. Through measurement investigations, we then obtain a better understanding of the non-verbal cues associated with structural events in order to model those structural events more effectively.

2.1 Multimodal Corpus Collection

Under NSF KDI award (Quek and et al.,), we collected a multimodal dialogue corpus. The corpus contains calibrated stereo video recordings, time-aligned word transcriptions, prosodic analyses, and hand positions tracked by a video tracking algorithm (Quek et al., 2002). To improve the speed of producing a corpus while maintaining its quality, we have investigated factors impacting the accuracy of the forced alignment of transcriptions to audio files (Chen et al., 2004a).

Meetings, in which several participants communicate with each other, play an important role in our daily life but increase the challenges to current information processing techniques. Understanding human multimodal communicative behavior, and how witting and unwitting visual displays (e.g., gesture, head orientation, gaze) relate to spoken content is critical to the analysis of meetings. These multimodal behaviors may reveal static and dynamic social structure of the meeting participants, the flow of topics being discussed, the control of floor of the meeting, and so on. For this purpose, we have been collecting a multimodal meeting corpus under the sponsorship of ARDA VACE II (Chen et al., 2005). In a room equipped with synchronized multichannel audio,video and motion-tracking recording devices, participants (from 5 to 8 civilian, military, or mixed) engage in planning exercises, such as managing rocket launch emergency, exploring a foreign weapon component, and collaborating to select awardees for fellowships. we have collected and continued to do multichannel time synchronized audio and video recordings. Using a series of audio and video processing techniques, we obtain the word transcriptions and prosodic features, as well as head, torso and hand $3D$ tracking traces from visual trackers and Vicon motion capture device. Figure 1 depicts our meeting corpus collection process.

2.2 Gesture Patterns during Speech Repairs

In the dynamic speech production process, speakers may make errors or totally change the content of what is being expressed. In either of these cases, speakers need refocus or revise what they are saying

and therefore speech repairs appear in overt speech. A typical speech repair contains a *reparandum*, an optional *editing phrase*, and a *correction*. Based on the relationship between the reparandum and the correction, speech repairs can be classified into three types: *repetitions*, *content replacements*, and *false starts*. Since utterance content has been modified in last two repair types, we call them content modification (**CM**) repairs. We carried out a measurement study (Chen et al., 2002) to identify patterns of gestures that co-occur with speech repairs that can be exploited by a multimodal processing system to more effectively process spontaneous speech. We observed that modification gestures (**MGs**), which exhibit a change in gesture state during speech repair, have a high correlation with content modification (**CM**) speech repairs, but rarely occur with content repetitions. This study does not only provide evidence that gesture and speech are tightly linked in production, but also provides evidence that gestures provide an important additional cue for identifying speech repairs and their types.

2.3 Incorporating Gesture in SU Detection

A sentence unit (SU) is defined as the complete expression of a speaker's thought or idea. It can be either a complete sentence or a semantically complete smaller unit. We have conducted an experiment that integrates lexical, prosodic and gestural cues in order to more effectively detect *sentence unit* boundaries in conversational dialog (Chen et al., 2004b).

As can be seen in Figure 2, our multimodal model combines lexical, prosodic, and gestural knowledge sources, with each knowledge source implemented as a separate model. A hidden event language model (LM) was trained to serve as lexical model ($P(W, E)$). Using a direct modeling approach (Shriberg and Stolcke, 2004), prosodic features were extracted using the SRI prosodic feature extraction tool[1] by collaborators at ICSI and then were used to train a CART decision tree as the prosodic model ($P(E|F)$). Similarly to the prosodic model, we computed gesture features directly from visual tracking measurements (Quek et al., 1999; Bryll et al., 2001): $3D$ hand position, Hold (a state when there is no hand motion beyond some adaptive threshold results), and Effort (analogous to the kinetic energy of hand movement). Using gestural features, we trained a CART tree to serve as the gestural model ($P(E|G)$). Finally, an HMM based model combination scheme was used to integrate predictions from individual models to obtain an overall SU prediction ($argmax(E|W, F, G)$). In our investigations, we found that gesture features complement the prosodic and lexical knowledge sources; by using all of the knowledge sources, the model is able to achieve the lowest overall detection error rate.

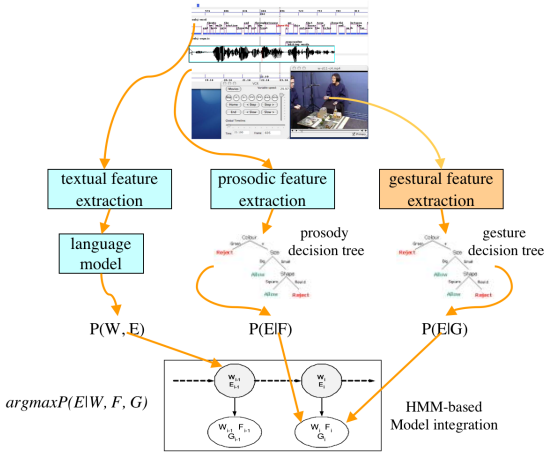


Figure 2: Data flow diagram of multimodal SU model using lexical, prosodic and gestural cues

2.4 Floor Control Investigation on Meetings

An underlying, auto-regulatory mechanism known as "floor control", allows participants communicate with each other coherently and smoothly. A person controlling the floor bears the burden of moving the discourse along. By increasing our understanding of floor control in meetings, there is a potential to impact two active research areas: human-like conversational agent design and automatic meeting analysis. We have recently investigated floor control in multi-party meetings (Chen et al., 2006). In particular, we analyzed patterns of speech (e.g., the use of *discourse markers*) and visual cues (e.g., eye gaze exchange, pointing gesture for next speaker) that are often involved in floor control changes. From this analysis, we identified some multimodal cues that will be helpful for predicting floor control events. Discourse markers are found to occur frequently at the beginning of a floor. During floor transitions, the

[1] A similar prosody feature extraction tool has been developed in our lab (Huang et al., 2006) using Praat.

previous holder often gazes at the next floor holder and vice verse. The well-known mutual gaze break pattern in dyadic conversations is also found in some meetings. A special participant, an active meeting manager, is found to play a role in floor transitions. Gesture cues are also found to play a role, especially with respect to floor capturing gestures.

3 Research Directions

In the next stage of my research, I will focus on integrating previous efforts into a complete multimodal model for structural event detection. In particular, I will improve current gesture feature extraction, and expand the non-verbal features to include both eye gaze and body posture. I will also investigate alternative integration architectures to the HMM shown in Figure 2. In my thesis, I hope to better understand the role that the non-verbal cues play in assisting structural event detection. My research is expected to support adding multimodal perception capabilities to current human communication systems that rely mostly on speech. I am also interested in investigating mutual impacts among the structural events. For example, we will study SUs and their relationship to floor control structure. Given progress in structural event detection in human communication, I also plan to utilize the detected structural events to further enhance meeting understanding. A particularly interesting task is to locate salient portions of a meeting from multimodal cues (Chen, 2005) to summarize it.

References

M. Argyle and M. Cook. 1976. *Gaze and Mutual Gaze*. Cambridge Univ. Press.

R. Bryll, F. Quek, and A. Esposito. 2001. Automatic hand hold detection in natural conversation. In *IEEE Workshop on Cues in Communication*, Kauai,Hawaii, Dec.

J. Cassell and M. Stone. 1999. Living Hand to Mouth: Psychological Theories about Speech and Gesture in Interactive Dialogue Systems. In *AAAI*.

L. Chen, M. Harper, and F. Quek. 2002. Gesture patterns during speech repairs. In *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, Pittsburg, PA, Oct.

L. Chen, Y. Liu, M. Harper, E. Maia, and S. McRoy. 2004a. Evaluating factors impacting the accuracy of forced alignments in a multimodal corpus. In *Proc. of Language Resource and Evaluation Conference*, Lisbon, Portugal, June.

L. Chen, Y. Liu, M. Harper, and E. Shriberg. 2004b. Multimodal model integration for sentence unit detection. In *Proc. of Int. Conf. on Multimodal Interface (ICMI)*, University Park, PA, Oct.

L. Chen, T.R. Rose, F. Parrill, X. Han, J. Tu, Z.Q. Huang, I. Kimbara, H. Welji, M. Harper, F. Quek, D. McNeill, S. Duncan, R. Tuttle, and T. Huang. 2005. VACE multimodal meeting corpus. In *Proceeding of MLMI 2005 Workshop*.

L. Chen, M. Harper, A. Franklin, T. R. Rose, I. Kimbara, Z. Q. Huang, and F. Quek. 2006. A multimodal analysis of floor control in meetings. In *Proc. of MLMI 06*, Washington, DC, USA, May.

L. Chen. 2005. Locating salient portions of meeting using multimodal cues. Research proposal submitted to AMI training program, Dec.

Z. Q. Huang, L. Chen, and M. Harper. 2006. An open source prosodic feature extraction tool. In *Proc. of Language Resource and Evaluation Conference*, May 2006.

Y. Liu, E. Shriberg, A. Stolcke, B. Peskin, J. Ang, Hillard D., M. Ostendorf, M. Tomalin, P. Woodland, and M. Harper. 2005. Structural Metadata Research in the EARS Program. In *Proc. of ICASSP*.

Y. Liu. 2004. *Structural Event Detection for Rich Transcription of Speech*. Ph.D. thesis, Purdue University.

D. McNeill. 1992. *Hand and Mind: What Gestures Reveal about Thought*. Univ. Chicago Press.

D. G. Novick. 2005. Models of gaze in multi-party discourse. In *Proc. of CHI 2005 Workshop on the Virtuality Continuum Revisted*, Portland OR, April 3.

F. Quek and et al. KDI: Cross-model Analysis Signal and Sense- Data and Computational Resources for Gesture, Speech and Gaze Research, http://vislab.cs.vt.edu/kdi.

F. Quek, R. Bryll, and X. F. Ma. 1999. A parallel algorighm for dynamic gesture tracking. In *ICCV Workshop on RATFG-RTS*, Gorfu,Greece.

F. Quek, D. McNeill, R. Bryll, S. Duncan, X. Ma, C. Kirbas, K. E. McCullough, and R. Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Trans. Comput.-Hum. Interact.*, 9(3):171–193.

E. Shriberg and A. Stolcke. 2004. Direct modeling of prosody: An overview of applications in automatic speech processing. In *International Conference on Speech Prosody*.

Semantic Back-Pointers from Gesture

Jacob Eisenstein
MIT Computer Science and Artificial Intelligence Laboratory
77 Massachusetts Ave, MA 02139
jacobe@csail.mit.edu

1 Introduction

Although the natural-language processing community has dedicated much of its focus to text, face-to-face spoken language is ubiquitous, and offers the potential for breakthrough applications in domains such as meetings, lectures, and presentations. Because spontaneous spoken language is typically more disfluent and less structured than written text, it may be critical to identify features from additional modalities that can aid in language understanding. However, due to the long-standing emphasis on text datasets, there has been relatively little work on non-textual features in unconstrained natural language (prosody being the most studied non-textual modality, e.g. (Shriberg et al., 2000)).

There are many non-verbal modalities that may contribute to face-to-face communication, including body posture, hand gesture, facial expression, prosody, and free-hand drawing. Hand gesture may be more expressive than any non-verbal modality besides drawing, since it serves as the foundation for sign languages in hearing-disabled communities. While non-deaf speakers rarely use any such systematized language as American Sign Language (ASL) while gesturing, the existence of ASL speaks to the potential of gesture for communicative expressivity.

Hand gesture relates to spoken language in several ways:

- Hand gesture communicates meaning. For example, (Kopp et al., 2006) describe a model of how hand gesture is used to convey spatial properties of its referents when speakers give navigational directions. This model both explains observed behavior of human speakers, and serves as the basis for an implemented embodied agent.
- Hand gesture communicates discourse structure. (Quek et al., 2002) and (McNeill, 1992) describe how the structure of discourse is mirrored by the the structure of the gestures, when speakers describe sequences of events in cartoon narratives.
- Hand gesture segments in unison with speech, suggesting possible applications to speech recognition and syntactic processing. (Morrel-Samuels and Krauss, 1992) show a strong correlation between the onset and duration of gestures, and their "lexical affiliates" – the phrase that is thought to relate semantically to the gesture. Also, (Chen et al., 2004) show that gesture features may improve sentence segmentation.

These examples are a subset of a broad literature on gesture that suggests that this modality could play an important role in improving the performance of NLP systems on spontaneous spoken language. However, the existence of significant relationships between gesture and speech does not prove that gesture will improve NLP; gesture features could be redundant with existing textual features, or they may be simply too noisy or speaker-dependant to be useful. To test this, my thesis research will identify specific, objective NLP tasks, and attempt to show that automatically-detected gestural features improve performance beyond what is attainable using textual features.

The relationship between gesture and meaning is particularly intriguing, since gesture seems to offer a unique, spatial representation of meaning to sup-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 215–218,
New York, June 2006.

plement verbal expression. However, the expression of meaning through gesture is likely to be highly variable and speaker dependent, as the set of possible mappings between meaning and gestural form is large, if not infinite. For this reason, I take the point of view that it is too difficult to attempt to decode individual gestures. A more feasible approach is to identify similarities between pairs or groups of gestures. If gestures do communicate semantics, then similar gestures should predict semantic similarity. Thus, gestures can help computers understand speech by providing a set of "back pointers" between moments that are semantically related. Using this model, my dissertation will explore measures of gesture similarity and applications of gesture similarity to NLP.

A set of semantic "back pointers" decoded from gestural features could be relevant to a number of NLP benchmark problems. I will investigate two: coreference resolution and disfluency detection. In coreference resolution, we seek to identify whether two noun phrases refer to the same semantic entity. A similarity in the gestural features observed during two different noun phrases might suggest a similarity in meaning. This problem has the advantage of permitting a quantitative evaluation of the relationship between gesture and semantics, without requiring the construction of a domain ontology.

Restarts are disfluencies that occur when a speaker begins an utterance, and then stops and starts over again. It is thought that the gesture may return to its state at the beginning of the utterance, providing a back-pointer to the restart insertion point (Esposito et al., 2001). If so, then a similar training procedure and set of gestural features can be used for both coreference resolution and restart correction. Both of these problems have objective, quantifiable success measures, and both may play an important role in bringing to spontaneous spoken language useful NLP applications such as summarization, segmentation, and question answering.

2 Current Status

My initial work involved hand annotation of gesture, using the system proposed in (McNeill, 1992). It was thought that hand annotation would identify relevant features to be detected by computer vision systems. However, in (Eisenstein and Davis, 2004), we found that the gesture phrase type (e.g., deictic, iconic, beat) could be predicted accurately by lexical information alone, without regard to hand movement. This suggests that this level of annotation inherently captures a synthesis of gesture and speech, rather than gesture alone. This conclusion was strengthened by (Eisenstein and Davis, 2005), where we found that hand-annotated gesture features correlate well with sentence boundaries, but that the gesture features were almost completely redundant with information in the lexical features, and did not improve overall performance.

The corpus used in my initial research was not suitable for automatic extraction of gesture features by computer vision, so a new corpus was gathered, using a better-defined experimental protocol and higher quality video and audio recording (Adler et al., 2004). An articulated upper body tracker, largely based on the work of (Deutscher et al., 2000), was used to identify hand and arm positions, using color and motion cues. All future work will be based on this new corpus, which contains six videos each from nine pairs of speakers. Each video is roughly two to three minutes in length.

Each speaker was presented with three different experimental conditions regarding how information in the corpus was to be presented: a) a pre-printed diagram was provided, b) the speaker was allowed to draw a diagram using a tracked marker, c) no presentational aids were allowed. The first condition was designed to be relevant to presentations involving pre-created presentation materials, such as Powerpoint slides. The second condition was intended to be similar to classroom lectures or design presentations. The third condition was aimed more at direct one-on-one interaction.

My preliminary work has involved data from the first condition, in which speakers gestured at preprinted diagrams. An empirical study on this part of the corpus has identified several gesture features that are relevant to coreference resolution (Eisenstein and Davis, 2006a). In particular, gesture similarity can be measured by hand position and the choice of the hand which makes the gesture; these similarities correlate with the likelihood of coreference. In addition, the likelihood of a gestural hold – where the hand rests in place for a period of

time – acts as a meta-feature, indicating that gestural cues are likely to be particularly important to disambiguate the meaning of the associated noun phrase. In (Eisenstein and Davis, 2006b), these features are combined with traditional textual features for coreference resolution, with encouraging results. The hand position gesture feature was found to be the fifth most informative feature by Chi-squared analysis, and the inclusion of gesture features yielded a statistically significant increase in performance over the textual features.

3 Future Directions

The work on coreference can be considered preliminary, because it is focused on a subset of our corpus in which speakers use pre-printed diagrams as an explanatory aide. This changes their gestures (Eisenstein and Davis, 2003), increasing the proportion of *deictic* gestures, in which hand position is the most important feature (McNeill, 1992). Hand position is assumed to be less useful in characterizing the similarity of *iconic* gestures, which express meaning through motion or handshape. Using the subsection of the corpus in which no explanatory aids were provided, I will investigate how to assess the similarity of such dynamic gestures, in the hope that coreference resolution can still benefit from gestural cues in this more general case.

Disfluency repair is another plausible domain in which gesture might improve performance. There are at least two ways in which gesture could be relevant to disfluency repair. Using the semantic backpointer model, restart repairs could be identified if there is a strong gestural similarity between the original start point and the restart. Alternatively, gesture could play a pragmatic function, if there are characteristic gestures that indicate restarts or other repairs. In one case, we are looking for a similarity between the disfluency and the repair point; in the other case, we are looking for similarities across all disfluencies, or across all repair points. It is hoped that this research will not only improve processing of spoken natural language, but also enhance our understanding of how speakers use gesture to structure their discourse.

4 Related Work

The bulk of research on multimodality in the NLP community relates to multimodal dialogue systems (e.g., (Johnston and Bangalore, 2000)). This research differs fundamentally from mine in that it addresses human-*computer* interaction, whereas I am studying human-*human* interaction. Multimodal dialogue systems tackle many interesting challenges, but the grammar, vocabulary, and recognized gestures are often pre-specified, and dialogue is controlled at least in part by the computer. In my data, all of these things are unconstrained.

Another important area of research is the generation of multimodal communication in animated agents (e.g., (Cassell et al., 2001; Kopp et al., 2006; Nakano et al., 2003)). While the models developed in these papers are interesting and often well-motivated by the psychological literature, it remains to be seen whether they are both broad and precise enough to apply to gesture recognition.

There is a substantial body of empirical work describing relationships between non-verbal and linguistic phenomena, much of which suggests that gesture could be used to improve the detection of such phenomena. (Quek et al., 2002) describe examples in which gesture correlates with topic shifts in the discourse structure, raising the possibility that topic segmentation and summarization could be aided by gesture features; Cassell et al. (2001) make a similar argument using body posture. (Nakano et al., 2003) describes how head gestures and eye gaze relate to turn taking and dialogue grounding. All of the studies listed in this paragraph identify relevant correlations between non-verbal communication and linguistic phenomena, but none construct a predictive system that uses the non-verbal modalities to improve performance beyond a text-only system.

Prosody has been shown to improve performance on several NLP problems, such as topic and sentence segmentation (e.g., (Shriberg et al., 2000; Kim et al., 2004)). The prosody literature demonstrates that non-verbal features can improve performance on a wide variety of NLP tasks. However, it also warns that performance is often quite sensitive, both to the representation of prosodic features, and how they are integrated with other linguistic features.

The literature on prosody would suggest parallels for gesture features, but little such work has been reported. (Chen et al., 2004) shows that gesture may improve sentence segmentation; however, in this study, the improvement afforded by gesture is not statistically significant, and evaluation was performed on a subset of their original corpus that was chosen to include only the three speakers who gestured most frequently. Still, this work provides a valuable starting point for the integration of gesture feature into NLP systems.

5 Summary

Spontaneous spoken language poses difficult problems for natural language processing, but these difficulties may be offset by the availability of additional communicative modalities. Using a model of hand gesture as providing a set of semantic back-pointers to previous utterances, I am exploring whether gesture can improve performance on quantitative NLP benchmark tasks. Preliminary results on coreference resolution are encouraging.

References

Aaron Adler, Jacob Eisenstein, Michael Oltmans, Lisa Guttentag, and Randall Davis. 2004. Building the design studio of the future. In *Making Pen-Based Interaction Intelligent and Natural*, pages 1–7, Menlo Park, California, October 21-24. AAAI Press.

Justine Cassell, Yukiko I. Nakano, Timothy W. Bickmore, Candace L. Sidner, and Charles Rich. 2001. Non-verbal cues for discourse structure. In *Proc. of ACL*, pages 106–115.

Lei Chen, Yang Liu, Mary P. Harper, and Elizabeth Shriberg. 2004. Multimodal model integration for sentence unit detection. In *Proceedings of International Conference on Multimodal Interfaces (ICMI'04)*. ACM Press.

Jonathan Deutscher, Andrew Blake, and Ian Reid. 2000. Articulated body motion capture by annealed particle filtering. In *IEEE Conference on Computer Vision and Pattern Recognition*, volume 2, pages 126–133.

Jacob Eisenstein and Randall Davis. 2003. Natural gesture in descriptive monologues. In *UIST'03 Supplemental Proceedings*, pages 69–70. ACM Press.

Jacob Eisenstein and Randall Davis. 2004. Visual and linguistic information in gesture classification. In *Proceedings of International Conference on Multimodal Interfaces(ICMI'04)*. ACM Press.

Jacob Eisenstein and Randall Davis. 2005. Gestural cues for sentence segmentation. Technical Report AIM-2005-014, MIT AI Memo.

Jacob Eisenstein and Randall Davis. 2006a. Gesture features for coreference resolution. In *Workshop on Multimodal Interaction and Related Machine Learning Algorithms*.

Jacob Eisenstein and Randall Davis. 2006b. Gesture improves coreference resolution. In *Proceedings of NAACL*.

Anna Esposito, Karl E. McCullough, and Francis Quek. 2001. Disfluencies in gesture: Gestural correlates to filled and unfilled speech pauses. In *Proceedings of IEEE Workshop on Cues in Communication*.

Michael Johnston and Srinivas Bangalore. 2000. Finite-state multimodal parsing and understanding,. In *Proceedings of COLING-2000*, pages 369–375.

Joungbum Kim, Sarah E. Schwarm, and Mari Osterdorf. 2004. Detecting structural metadata with decision trees and transformation-based learning. In *Proceedings of HLT-NAACL'04*. ACL Press.

Stefan Kopp, Paul Tepper, Kim Ferriman, and Justine Cassell. 2006. Trading spaces: How humans and humanoids use speech and gesture to give directions. *Spatial Cognition and Computation*, In preparation.

David McNeill. 1992. *Hand and Mind*. The University of Chicago Press.

P. Morrel-Samuels and R. M. Krauss. 1992. Word familiarity predicts temporal asynchrony of hand gestures and speech. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 18:615–623.

Yukiko Nakano, Gabe Reinstein, Tom Stocky, and Justine Cassell. 2003. Towards a model of face-to-face grounding. In *Proceedings of ACL'03*.

Francis Quek, David McNeill, Robert Bryll, Susan Duncan, Xin-Feng Ma, Cemil Kirbas, Karl E. McCullough, and Rashid Ansari. 2002. Multimodal human discourse: gesture and speech. *ACM Transactions on Computer-Human Interaction (TOCHI)*, pages 171–193.

Elizabeth Shriberg, Andreas Stolcke, Dilek Hakkani-Tur, and Gokhan Tur. 2000. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32.

Can the Internet help improve Machine Translation?

Ariadna Font Llitjós

Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA, 15213. USA
aria@cs.cmu.edu

Abstract

This paper summarizes a largely automated method that uses online post-editing feedback to automatically improve translation rules. As a starting point, bilingual speakers' local fixes are collected through an online Translation Correction Tool. Next, the Rule Refinement Module attacks the problem at its core and uses the local fixes to detect incorrect rules that need to be refined. Once the grammar and lexicon have been refined, the Machine Translation system not only produces the correct translation as fixed by the bilingual speaker, but is also able to generalize and correctly translates similar sentences. Thus, this work constitutes a novel approach to improving translation quality. Enhanced by the reaching power of the Internet, our approach becomes even more relevant to address the problem of how to automatically improve the quality of Machine Translation output.

1 Introduction

Achieving high translation quality remains the biggest challenge Machine Translation (MT) systems currently face. Researchers have explored a variety of methods to include user feedback in the MT loop. Similar to our approach, Phaholphinyo and colleagues (2005) proposed adding post-editing rules to their English-Thai MT system with the use of a post-editing tool. However, they use context sensitive pattern-matching rules, which make it impossible to fix errors involving missing words. Unlike our approach, in their system, the rules are created by experienced linguists and their approach requires a large corpus. They mention an experiment with 6,000 bilingual sentences but report no results due to data sparseness.

In general, most MT systems have failed to incorporate post-editing efforts beyond the addition of corrected translations to the parallel training data for SMT and EBMT or to a translation memory database.[1] Therefore, a largely automated method that uses online post-editing information to automatically improve translation rules constitutes a great advance in the field.

If an MT-produced translation is incorrect, a bilingual speaker can diagnose the presence of an error reliably using the online Translation Correction Tool (Font Llitjós and Carbonell, 2004). An example of an English-Spanish sentence pair generated by our MT system is "*Gaudí was a great artist - Gaudí era un artista grande*". Using the online tool, bilingual speakers modified the incorrect translation to obtain a correct one: "*Gaudí era un gran artista*".

Bilingual speakers, however, cannot be expected to diagnose which complex translation rules produced the error, and even less, determine how to improve those rules. One of the main goals of this research is to automate the Rule Refinement process based on just *error-locus* and possibly some *error-type* information from the bilingual speaker, relying on rule blame assignment and on regression testing to evaluate and measure the consequent improvement in MT accuracy. In this case, our Automatic Rule Refinement system can add the missing sense to the lexicon (great→gran) as

[1] For a more detailed discussion, see Font Llitjós and colleagues (2005a)

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 219–222,
New York, June 2006.

well as the special case rule for Spanish pre-nominal adjectives to the grammar.

With this system in place, we envision a modified version of the Translation Correction Tool as a game with a purpose, available online through a major web portal. This would allow bilingual speakers to correct MT input and get rewards for making good corrections, and compare their scores and speed with other users. For the MT community this means having a free and easy way to get MT output feedback and potentially improve their systems based on such feedback. Furthermore, a fully interactive system would be a great opportunity to show users that their corrections have a visible impact on technology, since they would see the effects their corrections have on other sentences. Last but not least, this new method is also expected to be particularly useful in resource-poor scenarios, such as the ones the Avenue project is devoted to (Font Llitjós et al., 2005b), where statistical systems are not an option and where there might be no experts with knowledge of the resource-poor language (Figure 1).

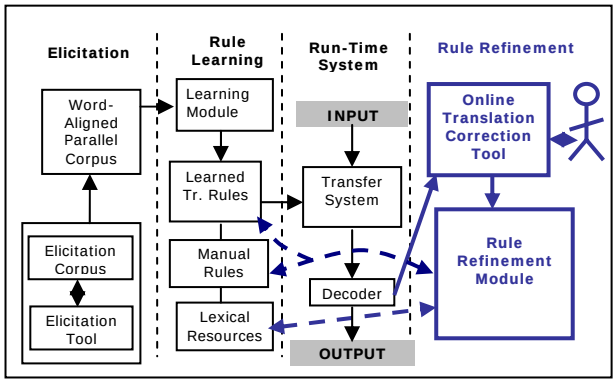


Figure 1. Simplified Avenue Architecture

2 Online Elicitation of MT Errors

The main challenge of the error elicitation part of this work is how to elicit minimal post-editing information from non-expert bilingual speakers. The Translation Correction Tool (TCTool) is a user-friendly online tool that allows users to add, delete and modify words and alignments, as well as to drag words around to change word order. A set of user studies was conducted to discover the right amount of error information that bilingual speakers can detect reliably when using the TCTool. These studies showed that simple error information can be elicited much more reliably (F1 0.89) than error type information (F1 0.72) (Font Llitjós and Carbonell, 2004). Most importantly, it became apparent that for our Rule Refinement purposes, the list of correction action(s) with information about error and correction words is sufficient.

Building on the example introduced above, Figure 2 shows the initial state of the TCTool, once the user has decided that the translation produced by the MT system is not correct.

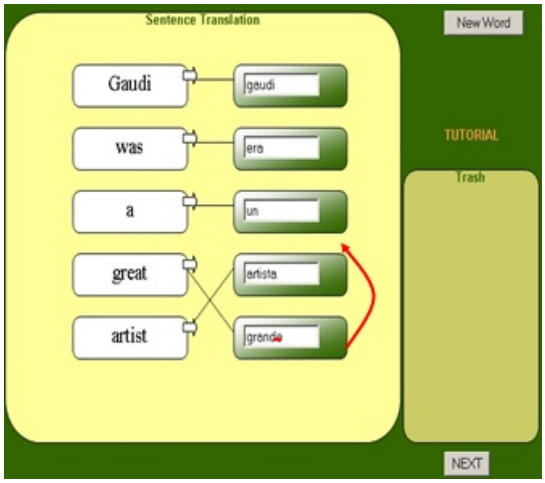


Figure 2. TCTool snapshot with initial translation pair

In this case, the bilingual speaker changed ‘grande’ to ‘gran’ and dragged ‘gran(de)’ in front of ‘artista’, effectively flipping the order of these two words. Figure 3 shows the state of the TCTool after the user corrections.

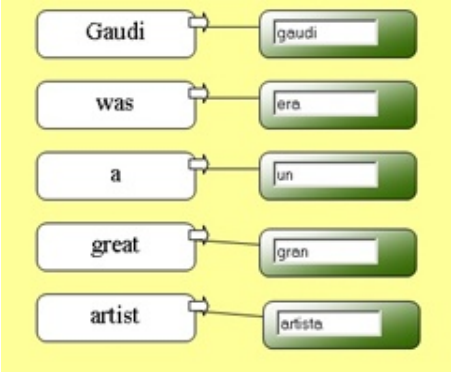


Figure 3. TCTool snapshot after user has corrected the translation

3 Extracting Error Information

User correction actions are registered into a log file. The Automatic Rule Refinement (RR) module extracts all the relevant information from the

TCTool log files and stores it into a Correction Instance. See Figure 4 for an example.

SL: *Gaudí was a great artist*
TL: *Gaudí era un artista grande*
AL: ((1,1),(2,2),(3,3),(4,5),(5,4))

Action 1: edit (*grande→ gran*)
Temp CTL: *Gaudi era un artista gran*

Action 2: change word order (*gran artista*)

CTL: *Gaudí era un gran artista*
AL: ((1,1),(2,2),(3,3),(4,4),(5,5))

Figure 4. A Correction Instance stores the source language sentence (SL), the target language sentence (TL) and the initial alignments (AL), as well as all the correction actions done by the user. It also provides the corrected translation (CTL) and final alignments.

The Rule Refinement (RR) module processes one action at a time. So in this approach, the order in which users correct a sentence does have an impact on the order in which refinements apply.

4 Lexical Refinements

After having stored all the relevant information from the log file, the Rule Refinement module starts processing the Correction Instance. In the example above, it first goes into the lexicon and, after double checking that there is no lexical entry for [great→gran], it proceeds to add one by duplicating the lexical entry for [great→grande]. Since these two lexical entries are identical at the feature level, the RR module postulates a new binary feature, say *feat1*[2], which serves the purpose of distinguishing between two words that are otherwise identical (according to our lexicon):

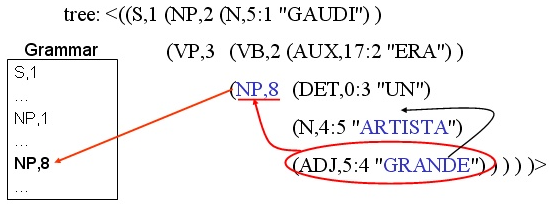
```
ADJ::ADJ |: [great] -> [grande]
((X1::Y1)
((x0 form) = great)
((y0 agr num) = sg)
((y0 agr gen) = masc)
((y0 feat1) = -))

ADJ::ADJ |: [great] -> [gran]
((X1::Y1)
((x0 form) = great)
((y0 agr num) = sg)
((y0 agr gen) = masc)
((y0 feat1) = +))
```

[2] A more mnemonic name for *feat1* would be *pre-nominal*.

5 Rule Refinements

Now the RR module moves on to process the next action in the Correction Instance and the first step is to look at the parse trace output by the MT system, so that the grammar rule responsible for the error can be identified:



At this point, the system extracts the relevant rule (NP,8) from the grammar, and has two options, either to make the required changes directly onto the original rule (REFINE) or to make a copy of the original rule and modify the copy (BIFURCATE). If the system has correctly applied the rule in the past (perhaps because users have evaluated the translation pair "*She saw a dangerous man – Ella vio un hombre peligroso*" as correct), then the RR module opts for the BIFURCATE operation. In this case, the RR module makes a copy of the original rule (NP,8) and then modifies the copy (NP,8') by flipping the order of the noun and adjective constituents, as indicated by the user. This rule needs to unify with 'gran' but not with 'grande', and so the RR module proceeds to add the constraint that the Spanish adjective (now y2) needs to have the *feat1* with value +:

```
{NP,8'}
NP::NP : [DET ADJ N] -> [DET ADJ N]
(  (X1::Y1)  (X2::Y2)  (X3::Y3)
((x0 def) = (x1 def))
(x0 = x3)
((y1 agr) = (y3 agr)) ; DET-N agreement
((y2 agr) = (y3 agr)) ; ADJ-N agreement
(y2 = x3)
((y2 feat1) =c + ))
```

These two refinements result in the MT system generating the desired translation, namely "*Gaudí era un gran artista*" and not the previous incorrect translation. But can the system also eliminate other incorrect translations automatically? In addition to generating the correct translation, we would also like the RR module to produce a refined grammar that is as tight as possible, given the data that is available. Since the system already has the information that "*un artista gran*" is not a correct se-

quence in Spanish, the grammar can be further refined to also rule out this incorrect translation. This can be done by restricting the application of the general rule (NP,8) to just post-nominal adjectives, like 'grande', which in this example are marked in the lexicon with (*feat1* = −).

6 Generalization power

The difference between this approach and mere post-editing is that the resulting refinements affect not only to the translation instance corrected by the user, but also to other similar sentences where the same error would manifest. After the refinements have been applied to the grammar in our example sentence, a sentence like "*Irina is a great friend*" will now correctly be translated as "*Irina es una gran amiga*", instead of "*Irina es una amiga grande*".

7 Evaluation

We plan to evaluate the RR module on its ability to improve coverage and overall translation quality. This requires identifying sensible evaluation metrics. Initial experiments have shown that both BLEU [Papineni et al., 2001] and METEOR [Lavie et al., 2004] can automatically distinguish between raw MT output and corrected MT output, even for a small set of sentences. In addition to the presence of the corrected translation in the lattice produced by the refined system, our evaluation metrics will also need to take into account whether the incorrect translation is now prevented from being generated and whether the lattice of alternative translations increased or decreased. A decrease of lattice size would mean that the refinement also made the grammar tighter, which is the desired effect.

8 Technical Challenges and Future Work

The Rule Refinement process is not invariable. It depends on the order in which refinement operations are applied. In batch mode, the RR module can rank Correction Instances (CI) in such a way as to maximize translation accuracy. Suppose that the first CI (CI1) triggers a bifurcation of a grammar rule, like the one we see in the example described in Section 5. After that, any CI that affects the same rule that got bifurcated, will only modify the original rule (NP,8) and not the copy (NP,8'). If the constraint that enforces determiner-noun agreement were missing from the original rule, say, the copy (NP,8') would not have that constraint added to it, and so another example with the pre-nominal adjective exhibiting that agreement error would be required (CI2: **Irina es un gran amiga*), before the system added the relevant constraint to NP,8'. However, if we can detect such rule dependencies before the refinement process, then we can try to find an optimal ranking, given the current set of CIs, which should result in higher translation accuracy, as measured on a test set.

Another interesting future direction is enhancing the Rule Refinement system to allow for further user interaction. In an interactive mode, the system can use Active Learning to produce minimal pairs to further investigate which refinement operations are more robust, treating the bilingual speaker as an oracle. We hope to explore the space between batch mode and a fully interactive system to discover the optimal setting which allows the system to only ask the user for further interaction when it cannot determine the appropriate refinement operation or when it would be impossible to correctly refine the grammar and the lexicon automatically.

References

Alon Lavie, Kenji Sagae and Shyamsundar Jayaraman. 2004. *The Significance of Recall in Automatic Metrics for MT Evaluation*. AMTA, Washington, DC.

Ariadna Font Llitjós, Jaime Carbonell and Alon Lavie. 2005a. *A Framework for Interactive and Automatic Refinement of Transfer-based Machine Translation*. EAMT, Budapest, Hungary.

Ariadna Font Llitjós, Roberto Aranovich and Lori Levin 2005b. *Building Machine translation systems for indigenous languages*. Second Conference on the Indigenous Languages of Latin America (CILLA II), Texas, USA.

Ariadna Font Llitjós and Jaime Carbonell. 2004. *The Translation Correction Tool: English-Spanish user studies*. LREC 04, Lisbon, Portugal.

Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. 2001. *BLEU: A Method for Automatic Evaluation of Machine Translation*. IBM Research Report RC22176 (W0109-022).

Sitthaa Phaholphinyo, Teerapong Modhiran, Nattapol Kritsuthikul and Thepchai Supnithi. 2005. *A Practical of Memory-based Approach for Improving Accuracy of MT*. MT Summit X. Phuket Island, Thailand.

Efficient Algorithms for Richer Formalisms: Parsing and Machine Translation

Liang Huang
Department of Computer and Information Science
University of Pennsylvania
lhuang3@cis.upenn.edu

My PhD research has been on the algorithmic and formal aspects of computational linguistics, esp. in the areas of parsing and machine translation. I am interested in developing efficient algorithms for formalisms with rich expressive power, so that we can have a better modeling of human languages without sacrificing efficiency. In doing so, I hope to help integrating more linguistic and structural knowledge with modern statistical techniques, and in particular, for syntax-based machine translation (MT) systems.

Among other projects, I have been working on k-best parsing, synchronous binarization, and syntax-directed translation.

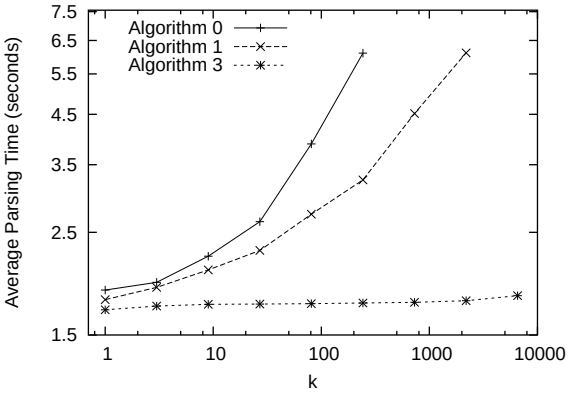


Figure 1: Average parsing speed on the Section 23 of Penn Treebank (Algorithms 0, 1, and 3, log-log).

1 k-best Parsing and Hypergraphs

NLP systems are often cascades of several modules, e.g., part-of-speech tagging, then syntactic parsing, and finally semantic interpretation. It is often the case that the 1-best output from one module is *not* always optimal for the next module. So one might want to postpone some disambiguation by propagating k-best lists (instead of 1-best solutions) to subsequent phases, as in joint parsing and semantic role-labeling (Gildea and Jurafsky, 2002). This is also true for *reranking* and *discriminative training*, where the k-best list of candidates serves as an approximation of the full set (Collins, 2000; Och, 2003; McDonald et al., 2005). In this way we can optimize some complicated objective function on the k-best set, rather than on the full search space which is usually exponentially large.

Previous algorithms for k-best parsing (Collins, 2000; Charniak and Johnson, 2005) are either sub-optimal or slow and rely significantly on pruning techniques to make them tractable. So I co-developed several fast and exact algorithms for k-best parsing in the general framework of directed monotonic hypergraphs (Huang and Chiang, 2005). This formulation extends and refines Klein and Manning's work (2001) by introducing *monotonic weight functions*, which is closely related to the optimal subproblem property in dynamic programming.

We first generalize the classical 1-best Viterbi algorithm to hypergraphs, and then present four k-best algorithms, each improving its predessor by delaying more work until necessary. The final one, Algorithm 3, starts with a normal 1-best search for each vertex (or *item*, as in deductive frameworks), and then works backwards from the target vertex (final item) for its 2nd, 3rd, . . ., kth best derivations, calling itself recursively only on demand, being the laziest of the four algorithms. When tested on top of two state-of-the-art systems, the Collins/Bikel parser (Bikel, 2004) and Chiang's CKY-based Hiero decoder (Chiang, 2005), this algorithm is shown to have very little overhead even for quite large k (say, 10^6) (See Fig. 1 for experiments on Bikel parser).

These algorithms have been re-implemented by other researchers in the field, including Eugene Charniak for his n-best parser, Ryan McDonald for his dependency parser (McDonald et al., 2005), Microsoft Research NLP group (Simon Corston-Oliver and Kevin Duh, p.c.) for a similar model, Jonathan Graehl for the ISI syntax-based MT decoder, David A. Smith for the Dyna language (Eisner et al., 2005),

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 223–226,
New York, June 2006.

and Jonathan May for ISI's tree automata package Tiburon. All of these experiments confirmed the findings in our work.

2 Synchronous Binarization for MT

Machine Translation has made very good progress in recent times, especially, the so-called "phrase-based" statistical systems (Och and Ney, 2004). In order to take a substantial next-step it will be necessary to incorporate several aspects of syntax. Many researchers have explored syntax-based methods, for instance, Wu (1996) and Chiang (2005) both uses binary-branching synchronous context-free grammars (SCFGs). However, to be more expressive and flexible, it is often easier to start with a general SCFG or tree-transducer (Galley et al., 2004). In this case, *binarization* of the input grammar is required for the use of the CKY algorithm (in order to get cubic-time complexity), just as we convert a CFG into the Chomsky Normal Form (CNF) for monolingual parsing. For synchronous grammars, however, different binarization schemes may result in very different-looking chart items that greatly affect decoding efficiency. For example, consider the following SCFG rule:

(1) $\text{S} \rightarrow \text{NP}^{(1)}\ \text{VP}^{(2)}\ \text{PP}^{(3)},\ \text{NP}^{(1)}\ \text{PP}^{(3)}\ \text{VP}^{(2)}$

We can binarize it either left-to-right or right-to-left:

$$\begin{array}{rl} \text{S} \rightarrow & V_{\text{NP-PP}}\ \text{VP} \\ V_{\text{NP-PP}} \rightarrow & \text{NP PP} \end{array} \quad \text{or} \quad \begin{array}{rl} \text{S} \rightarrow & \text{NP}\ V_{\text{PP-VP}} \\ V_{\text{PP-VP}} \rightarrow & \text{PP VP} \end{array}$$

The intermediate symbols (e.g. $V_{\text{PP-VP}}$) are called *virtual nonterminals*. We would certainly prefer the right-to-left binarization because the virtual nonterminal has consecutive span (see Fig. 2). The left-to-right binarization causes discontinuities on the target side, which results in an exponential time complexity when decoding with an integrated n-gram model.

We develop this intuition into a technique called *synchronous binarization* (Zhang et al., 2006) which binarizes a synchronous production or tree-tranduction rule on both source and target sides *simultaneously*. It essentially converts an SCFG into an equivalent ITG (the synchronous extension of CNF) if possible. We reduce this problem to the binarization of the permutation of nonterminal symbols between the source and target sides of a synchronous rule and devise a linear-time algorithm for it. Experiments show that the resulting rule set significantly improves the speed and accuracy over monolingual binarization (see Table 1) in a state-of-the-art syntax-based machine translation system (Galley et al., 2004). We also propose another trick (hook) for further speeding up the decoding with integrated n-gram models (Huang et al., 2005).

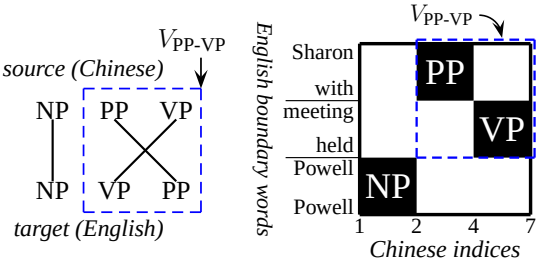


Figure 2: The alignment pattern (left) and alignment matrix (right) of the SCFG rule.

system	BLEU
monolingual binarization	36.25
synchronous binarization	*38.44*

Table 1: Synchronous vs. monolingual binarization in terms of translation quality (BLEU score).

3 Syntax-Directed Translation

Syntax-directed translation was originally proposed for compiling programming languages (Irons, 1961; Lewis and Stearns, 1968), where the source program is parsed into a syntax-tree that guides the generation of the object code. These translations have been formalized as a synchronous context-free grammar (SCFG) that generates two languages simultaneously (Aho and Ullman, 1972), and equivalently, as a top-down tree-to-string transducer (Gécseg and Steinby, 1984). We adapt this syntax-directed transduction process to statistical MT by applying stochastic operations at each node of the source-language parse-tree and searching for the best derivation (a sequence of translation steps) that converts the whole tree into some target-language string (Huang et al., 2006).

3.1 Extended Domain of Locality

From a modeling perspective, however, the structural divergence across languages results in non-isomorphic parse-trees that are not captured by

SCFGs. For example, the S(VO) structure in English is translated into a VSO order in Arabic, an instance of *complex re-ordering* (Fig. 4).

To alleviate this problem, grammars with richer expressive power have been proposed which can grab larger fragments of the tree. Following Galley et al. (2004), we use an extended tree-to-string transducer (**xRs**) with multi-level left-hand-side (LHS) trees.[1] Since the right-hand-side (RHS) string can be viewed as a flat one-level tree with the same non-terminal root from LHS (Fig. 4), this framework is closely related to STSGs in having extended domain of locality on the source-side except for remaining a CFG on the target-side. These rules can be learned from a parallel corpus using English parse-trees, Chinese strings, and word alignment (Galley et al., 2004).

3.2 A Running Example

Consider the following English sentence and its Chinese translation (note the reordering in the passive construction):

(2) the gunman was killed by the police .

qiangshou bei jingfang jibi 。
[gunman] [passive] [police] [killed] .

Figure 3 shows how the translator works. The English sentence (a) is first parsed into the tree in (b), which is then recursively converted into the Chinese string in (e) through five steps. First, at the root node, we apply the rule r_1 which preserves the top-level word-order and translates the English period into its Chinese counterpart:

(r_1) S (x_1:NP-C x_2:VP PUNC (.)) $\rightarrow x_1\ x_2$ 。

Then, the rule r_2 grabs the whole sub-tree for "the gunman" and translates it as a phrase:

(r_2) NP-C (DT (the) NN (gunman)) $\rightarrow$ *qiangshou*

Now we get a "partial Chinese, partial English" sentence "*qiangshou* VP 。" as shown in Fig. 3 (c). Our recursion goes on to translate the VP sub-tree. Here we use the rule r_3 for the passive construction:

[1] we will use LHS and source-side interchangeably (so are RHS and target-side). In accordance with our experiments, we also use English and Chinese as the source and target languages, opposite to the Foreign-to-English convention of Brown et al. (1993).

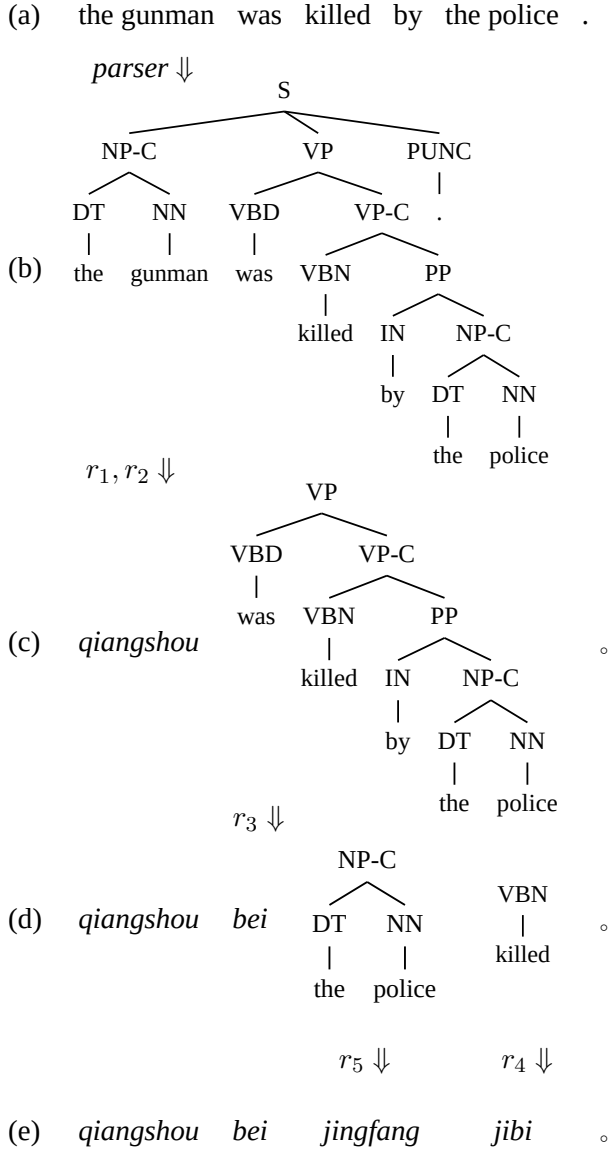


Figure 3: A synatx-directed translation process.

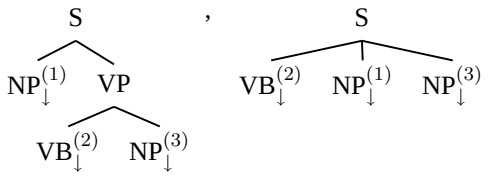
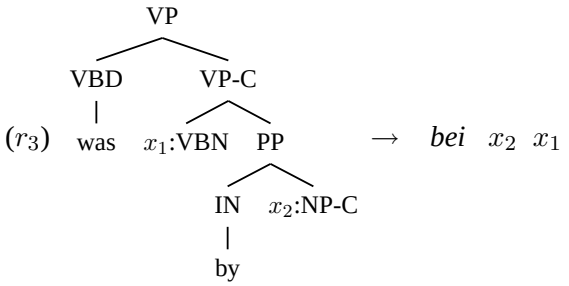


Figure 4: An example of complex re-ordering.



which captures the fact that the agent (NP-C, "the police") and the verb (VBN, "killed") are always inverted between English and Chinese in a passive voice. Finally, we apply rules r_4 and r_5 which perform phrasal translations for the two remaining subtrees in (d), respectively, and get the completed Chinese string in (e).

3.3 Translation Algorithm

Given a fixed parse-tree τ^*, the search for the best derivation (as a sequence of conversion steps) can be done by a simple top-down traversal (or depth-first search) from the root of the tree. With memoizationm, we get a dynamic programming algorithm that is guaranteed to run in $O(n)$ time where n is the length of the input string, since the size of the parse-tree is proportional to n. Similar algorithms have also been proposed for dependency-based translation (Lin, 2004; Ding and Palmer, 2005).

I am currently performing large-scale experiments on English-to-Chinese translation using the **xRs** rules. We are not doing the usual direction of Chinese-to-English partly due to the lack of a sufficiently good Chinese parser. Initial results show promising translation quality (in terms of BLEU scores) and fast translation speed.

References

Alfred V. Aho and Jeffrey D. Ullman. 1972. *The Theory of Parsing, Translation, and Compiling*, volume I: Parsing of *Series in Automatic Computation*. Prentice Hall, Englewood Cliffs, New Jersey.

Daniel M. Bikel. 2004. Intricacies of Collins' parsing model. *Computational Linguistics*, 30(4):479–511, December.

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19:263–311.

Eugene Charniak and Mark Johnson. 2005. Coarse-to-fi negrained n-best parsing and discriminative reranking. In *Proceedings of the 43rd ACL*.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proc. of the 43rd ACL*.

Michael Collins. 2000. Discriminative reranking for natural language parsing. In *Proceedings of ICML*, pages 175–182.

Yuan Ding and Martha Palmer. 2005. Machine translation using probablisitic synchronous dependency insertion grammars. In *Proceedings of the 43rd ACL*.

Jason Eisner, Eric Goldlust, and Noah A. Smith. 2005. Compiling comp ling: Weighted dynamic programming and the dyna language. In *Proceedings of HLT-EMNLP*.

Michel Galley, Mark Hopkins, Kevin Knight, and Daniel Marcu. 2004. What's in a translation rule? In *HLT-NAACL*.

F. Gécseg and M. Steinby. 1984. *Tree Automata*. Akadémiai Kiadó, Budapest.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

Liang Huang and David Chiang. 2005. Better k-best Parsing. In *Proceedings of 9th International Workshop of Parsing Technologies (IWPT)*.

Liang Huang, Hao Zhang, and Daniel Gildea. 2005. Machine translation as lexicalized parsing with hooks. In *Proceedings of 9th International Workshop of Parsing Technologies (IWPT)*.

Liang Huang, Kevin Knight, and Aravind Joshi. 2006. Syntax-directed translation with extended domain of locality. In submission.

E. T. Irons. 1961. A syntax-directed compiler for ALGOL 60. *Comm. ACM*, 4(1):51–55.

Dan Klein and Christopher D. Manning. 2001. Parsing and Hypergraphs. In *Proceedings of the Seventh International Workshop on Parsing Technologies (IWPT-2001), 17-19 October 2001, Beijing, China*.

P. M. Lewis and R. E. Stearns. 1968. Syntax-directed transduction. *Journal of the ACM*, 15(3):465–488.

Dekang Lin. 2004. A path-based transfer model for machine translation. In *Proceedings of the 20th COLING*.

Ryan McDonald, Koby Crammer, and Fernando Pereira. 2005. Online large-margin training of dependency parsers. In *Proceedings of the 43rd ACL*.

F. J. Och and H. Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449.

Franz Och. 2003. Minimum error rate training for statistical machine translation. In *Proc. of ACL*.

Dekai Wu. 1996. A polynomial-time algorithm for statistical machine translation. In *Proceedings of the 34th ACL*.

Hao Zhang, Liang Huang, Daniel Gildea, and Kevin Knight. 2006. Synchronous binarization for machine translation. In *Proceedings of HLT-NAACL*.

Identifying Perspectives at the Document and Sentence Levels Using Statistical Models

Wei-Hao Lin*
Language Technologies Institute
Carnegie Mellon University
Pittsburgh, PA 15213 U.S.A.
whlin@cs.cmu.edu

Abstract

In this paper we investigate the problem of identifying the perspective from which a document was written. By perspective we mean a point of view, for example, from the perspective of Democrats or Republicans. Can computers learn to identify the perspective of a document? Furthermore, can computers identify which sentences in a document strongly convey a particular perspective? We develop statistical models to capture how perspectives are expressed at the document and sentence levels, and evaluate the proposed models on a collection of articles on the Israeli-Palestinian conflict. The results show that the statistical models can successfully learn how perspectives are reflected in word usage and identify the perspective of a document with very high accuracy.

1 Introduction

In this paper we investigate the problem of automatically identifying the *perspective* from which a document was written. By perspective, we mean "subjective evaluation of relative significance, a point-of-view." For example, documents about the Palestinian-Israeli conflict may appear to be about the same topic, but reveal different perspectives:

(1) The inadvertent killing by Israeli forces of Palestinian civilians – usually in the course of shooting at Palestinian terrorists – is considered no different at the moral and ethical level than the deliberate targeting of Israeli civilians by Palestinian suicide bombers.

(2) In the first weeks of the Intifada, for example, Palestinian public protests and civilian demonstrations were answered brutally by Israel, which killed tens of unarmed protesters.

Example 1 is written from a Israeli perspective; Example 2 is written from a Palestinian perspective . We aim to address a research question: can computers learn to identify the perspective of a document given a training corpus of documents that are written from different perspectives?

When an issue is discussed from different perspectives, not every sentence in a document strongly reflects the perspective the author possesses. For example, the following sentences are written by one Palestinian and one Israeli:

(3) The Rhodes agreements of 1949 set them as the ceasefire lines between Israel and the Arab states.

(4) The green line was drawn up at the Rhodes Armistice talks in 1948-49.

Example 3 and 4 both factually introduce the background of the issue of the "green line" without expressing explicit perspectives. Can computers automatically discriminate between sentences that strongly express a perspective and sentences that only reflect shared background information?

*This is joint work with Theresa Wilson, Janyce Wiebe, and Alexander Hauptmann, and supported by the Advanced Research and Development Activity (ARDA) under contract number NBCHC040037.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 227–230,
New York, June 2006.

A system that can automatically identify the perspective from which a document written will be a highly desirable tool for people analyzing huge collections of documents from different perspectives. An intelligence analyst regularly monitors the positions that foreign countries take on political and diplomatic issues. A media analyst frequently surveys broadcast news, newspapers, and web blogs for different viewpoints. What these analysts need in common is that they would like to find evidence of strong statements of differing perspectives, while ignoring statements without strong perspectives as less interesting.

In this paper we approach the problem of learning perspectives in a statistical learning framework. We develop statistical models to learn how perspectives are reflected in word usage, and evaluate the models by measuring how accurately they can predict the perspectives of unseen documents. Lacking annotation on how strongly individual sentences convey a particular perspective in our corpus poses a challenge on learning sentence-level perspectives. We propose a novel statistical model, Latent Sentence Perspective Model, to address the problem.

2 Related Work

Identifying the perspective from which a document is written is a subtask in the growing area of automatic opinion recognition and extraction. Subjective language is used to express opinions, emotions, and sentiments. So far research in automatic opinion recognition has primarily addressed learning subjective language (Wiebe et al., 2004; Riloff et al., 2003; Riloff and Wiebe, 2003), identifying opinionated documents (Yu and Hatzivassiloglou, 2003) and sentences (Yu and Hatzivassiloglou, 2003; Riloff et al., 2003; Riloff and Wiebe, 2003), and discriminating between positive and negative language (Yu and Hatzivassiloglou, 2003; Turney and Littman, 2003; Pang et al., 2002; Dave et al., 2003; Nasukawa and Yi, 2003; Morinaga et al., 2002).

Although by its very nature we expect much of the language of presenting a perspective or point-of-view to be subjective, labeling a document or a sentence as subjective is not enough to identify the perspective from which it is written. Moreover, the ideology and beliefs authors possess are often expressed in ways more than conspicuous positive or negative language toward specific targets.

3 Corpus

Our corpus consists of articles published on the `bitterlemons` website[1]. The website is set up to "contribute to mutual understanding [between Palestinians and Israels] through the open exchange of ideas". Every week an issue about Israeli-Palestinian conflict is selected for discussion, for example, "Disengagement: unilateral or coordinated?", and a Palestinian editor and an Israeli editor contribute a article addressing the issue. In addition, the Israeli and Palestinian editors invite or interview one Israeli and one Palestinian to express their views, resulting in a total of four articles in a weekly edition.

We evaluate the subjectivity of each sentence using the patterns automatically extracted from foreign news documents (Riloff and Wiebe, 2003), and find that 65.6% of Palestinian sentences and 66.2% of Israeli sentences are classified as subjective. The high but almost equivalent percentages of subjective sentences from two perspectives supports our observation in Section 2 that perspective is largely expressed in subjective language but subjectivity ratio is not necessarily indicative of the perspective of a document.

4 Statistical Modeling of Perspectives

We approach the problem of learning perspectives in a statistical learning framework. Denote a training corpus as pairs of documents W_n and their perspectives labels D_n, $n = 1, \ldots, N$, N is the total number of documents in the corpus. Given a new document $\tilde{W}$ with a unknown document perspective $\tilde{D}$, identifying its perspective is to calculate the following conditional probability,

$$P(\tilde{D}|\tilde{W}, \{D_n, W_n\}_{n=1}^{N}) \quad (5)$$

We are interested in how strongly each sentence in the document convey perspective. Denote the intensity of the m-th sentence of the n-th document as a binary random variable $S_{m,n}$, $m = 1, \ldots, M_n$, M_n is the total number of sentences of the n-th document. Evaluating how strongly a sentence conveys

[1] `http://www.bitterlemons.org`

a particular perspective is to calculate the following conditional probability,

$$P(S_{m,n}|\{D_n, W_n\}_{n=1}^{N}) \quad (6)$$

4.1 Document Perspective Models

The process of generating documents from a particular perspective is modeled as follows,

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_\pi, \beta_\pi) \\ \theta &\sim \text{Dirichlet}(\alpha_\theta) \\ D_n &\sim \text{Binomial}(1, \pi) \\ W_n &\sim \text{Multinomial}(L_n, \theta_d) \end{aligned}$$

The model is known as naïve Bayes models (NB), which has been widely used for NLP tasks such as text categorization (Lewis, 1998). To calculate (5) under NB in a full Bayesian manner is, however, complicated, and alternatively we employ Markov Chain Monte Carlo (MCMC) methods to simulate samples from the posterior distributions.

4.2 Latent Sentence Perspective Models

We introduce a new binary random variables, S, to model how strongly a perspective is expressed at the sentence level. The value of S is either s^1 or s^0, where s^1 means the sentence is written strongly from a perspective, and s^0 is not. The whole generative process is modeled as follows,

$$\begin{aligned} \pi &\sim \text{Beta}(\alpha_\pi, \beta_\pi) \\ \tau &\sim \text{Beta}(\alpha_\tau, \beta_\tau) \\ \theta &\sim \text{Dirichlet}(\alpha_\theta) \\ D_n &\sim \text{Binomial}(1, \pi) \\ S_{m,n} &\sim \text{Binomial}(1, \tau) \\ W_{m,n} &\sim \text{Multinomial}(L_{m,n}, \theta) \end{aligned}$$

π and θ carry the same semantics as those in NB. S is naturally modeled as a binary variable, where τ is the parameter of S and represents how likely a perspective is strongly expressed at the sentence given on the overall document perspective. We call this model **Latent Sentence Perspective Models** (LSPM), because S is never directly observed in either training or testing documents and need to be inferred. To calculate (6) under LSPM is difficult. We again resort to MCMC methods to simulate samples from the posterior distributions.

5 Experiments

5.1 Identifying Perspectives at the Document Level

To objectively evaluate how well naïve Bayes models (NB) learn to identify perspectives expressed at the document level, we train NB against on the `bitterlemons` corpus, and evaluate how accurately NB predicts the perspective of a unseen document as either Palestinian or Israeli in ten-fold cross-validation manner. The average classification accuracy over 10 folds is reported. We compare three different models, including NB with two different inference methods and Support Vector Machines (SVM) (Cristianini and Shawe-Taylor, 2000). NB-B uses full Bayesian inference and NB-M uses Maximum a posteriori (MAP).

Model	Data Set	Accuracy	Reduction
Baseline		0.5	
SVM	Editors	0.9724	
NB-M	Editors	0.9895	61%
NB-B	Editors	0.9909	67%
SVM	Guests	0.8621	
NB-M	Guests	0.8789	12%
NB-B	Guests	0.8859	17%

Table 1: Results of Identifying Perspectives at the Document Level

The results in Table 1 show that both NB and SVM perform surprisingly well on both Editors and Guests subsets of the `bitterlemons` corpus. We also see that NBs further reduce classification errors even though SVM already achieves high accuracy. By considering the full posterior distribution NB-B further improves on NB-M, which performs only point estimation. The results strongly suggest that the word choices made by authors, either consciously or subconsciously, reflect much of their political perspectives.

5.2 Identifying Perspectives at the Sentence Level

In addition to identify the perspectives of a document, we are interested in which sentences in the document strongly convey perspectives. Although the posterior probability that a sentence

covey strongly perspectives in (6) is of our interest, we can not directly evaluate their quality due to the lack of golden truth at the sentence level. Alternatively we evaluate how accurately LSPM predicts the perspective of a document, in the same way of evaluating SVM and NB in the previous section. If LSPM does not achieve similar identification accuracy after modeling sentence-level information, we will doubt the quality of predictions on how strongly a sentence convey perspective made by LSPM.

Model	Training	Testing	Accuracy
Baseline			0.5
NB-M	Guest	Editor	0.9327
NB-B	Guest	Editor	0.9346
LSPM	Guest	Editor	0.9493
NB-M	Editors	Guests	0.8485
NB-B	Editors	Guests	0.8585
LSPM	Guest	Editor	0.8699

Table 2: Results of Perspective Identification at the Sentence Level

The experimental results in Table 2 show that the LSPM achieves similarly or even slightly better accuracy than those of NBs, which is very encouraging and suggests that the proposed LSPM closely match how perspectives are expressed at the document and sentence levels. If one does not explicitly model the uncertainty at the sentence level, one can train NB directly against the sentences to classify a sentence into Palestinian or Israeli perspective. We obtain the accuracy of 0.7529, which is much lower than the accuracy previously achieved at the document level. Therefore identifying perspective at the sentence level is much harder than at that the document level, and the high accuracy of identifying document-level perspectives suggests that LPSM closely captures the perspectives expressed at the document and sentence levels, given individual sentences are very short and much less informative about overall perspective.

6 Summary of Contributions

In this paper we study the problem of learning to identify the perspective from which a text was written at the document and sentence levels. We show that perspectives are expressed in word usage, and statistical learning algorithms such as SVM and naïve Bayes models can successfully uncover the word patterns chosen by authors from different perspectives. Furthermore, we develop a novel statistical model to infer how strongly a sentence convey perspective without any labels. By introducing latent variables, Latent Sentence Perspective Models are shown to capture well how perspectives are reflected at the document and sentence levels. The proposed statistical models can help analysts sift through a large collection of documents written from different perspectives. The unique sentence-level perspective modeling can automatically identify sentences that are strongly representative of the perspective of interest, and we plan to manually evaluate their quality in the future work.

References

Nello Cristianini and John Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press.

Kushal Dave, Steve Lawrence, and David M. Pennock. 2003. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th International World Wide Web Conference (WWW2003)*.

David D. Lewis. 1998. Naive (Bayes) at forty: The independence assumption in information retrieval. In *Proceedings of the European Conference on Machine Learning (ECML)*.

S. Morinaga, K. Yamanishi, K. Tateishi, and T. Fukushima. 2002. Mining product reputations on the web. In *Proceedings of the 2002 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

T. Nasukawa and J. Yi. 2003. Sentiment analysis: Capturing favorability using natural language processing. In *Proceedings of the 2nd International Conference on Knowledge Capture (K-CAP 2003).*

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2002).*

Ellen Riloff and Janyce Wiebe. 2003. Learning extraction patterns for subjective expressions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003)*.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the 7th Conference on Natural Language Learning (CoNLL-2003).*

Peter Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346.

Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. 2004. Learning subjective language. *Computational Linguistics*, 30(3).

Hong Yu and Vasileios Hatzivassiloglou. 2003. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-2003).*

Detecting Emotion in Speech: Experiments in Three Domains

Jackson Liscombe
Columbia University
jaxin@cs.columbia.edu

Abstract

The goal of my proposed dissertation work is to help answer two fundamental questions: (1) How is emotion communicated in speech? and (2) Does emotion modeling improve spoken dialogue applications? In this paper I describe feature extraction and emotion classification experiments I have conducted and plan to conduct on three different domains: EPSaT, HMIHY, and ITSpoke. In addition, I plan to implement emotion modeling capabilities into ITSpoke and evaluate the effectiveness of doing so.

1 Introduction

The focus of my work is the expression of emotion in human speech. As normally-functioning people, we are each capable of vocally expressing and aurally recognizing the emotions of others. How often have you been put off by the "tone in someone's voice" or tickled others with the humorous telling of a good story? Though we as everyday people are intimately familiar with emotion, we as scientists do not actually know precisely how it is that emotion is conveyed in human speech. This is of special concern to us as engineers of natural language technology; in particular, spoken dialogue systems. Spoken dialogue systems enable users to interact with computer systems via natural dialogue, as they would with human agents. In my view, a current deficiency of state-of-the-art spoken dialogue systems is that the emotional state of the user is not modeled. This results in non-human-like and even inappropriate behavior on the part of the spoken dialogue system.

There are two central questions I would like to at lest partially answer with my dissertation research: (1) How is emotion communicated in speech? and (2) Does emotion modeling improve spoken dialogue applications? In an attempt to answer the first question, I have adopted the research paradigm of extracting features that characterize emotional speech and applying machine learning algorithms to determine the prediction accuracy of each feature. With regard to the second research question, I plan to implement an emotion modeler – one that detects and responds to uncertainty and frustration – into an Intelligent Tutoring System.

2 Completed Work

This section describes my current research on emotion classification in three domains and forms the foundation of my dissertation. For each domain, I have adopted an experimental design wherein each utterance in a corpus is annotated with one or more emotion labels, features are extracted from these utterances, and machine learning experiments are run to determine emotion prediction accuracy.

2.1 EPSaT

The publicly-available Emotional Prosody Speech and Transcription corpus[1] (EPSaT) comprises recordings of professional actors reading short (four syllables each) dates and numbers (*e.g.*, 'two-thousand-four') with different emotional states. I chose a subset of 44 utterances from 4 speakers (2 male, 2 female) from this corpus and conducted a web-based survey to subjectively label each utterance for each of 10 emotions, divided evenly for valence. These emotions included the positive emotion categories: *confident*, *encouraging*, *friendly*, *happy*, *interested*; and the negative emotion categories: *angry*, *anxious*, *bored*, *frustrated*, *sad*.

Several features were extracted from each utterance in this corpus, each one designed to capture emotional content. Global acoustic-prosodic information – *e.g.*, speaking rate and minimum, maximum, and mean pitch and intensity – has been well known since the 1960s and 1970s

[1] LDC Catalog No.: LDC2002S28.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 231–234,
New York, June 2006.

to convey emotion to some extent (*e.g.*, (Davitz, 1964; Scherer et al., 1972)). In addition to these features, I also included linguistically meaningful prosodic information in the form of ToBI labels (Beckman et al., 2005), as well as the spectral tilt of the vowel in each utterance bearing the nuclear pitch accent.

In order to evaluate the predictive power of each feature extracted from the EPSaT utterances, I ran machine learning experiments using RIPPER, a rule-learning algorithm. The EPSaT corpus was divided into training (90%) and testing (10%) sets. A binary classification scheme was adopted based on the observed ranking distributions from the perception survey: "*not at all*" was considered to be the absence of emotion *x*; all other ranks was recorded as the presence of emotion *x*. Performance accuracy varied with respect to emotion, but on average I observed 75% prediction accuracy for any given emotion, representing an average 22% improvement over chance performance. The most predictive included the global acoustic-prosodic features, but interesting novel findings emerged as well; most notably, significant correlation was observed between negative emotions and pitch contours ending in a plateau boundary tone, whereas positive emotions correlated with the standard declarative phrasal ending (in ToBI, these would be labeled as /H-L%/ and /L-L%/, respectively). Further discussion of such findings can be found in (Liscombe et al., 2003).

2.2 HMIHY

"How May I Help You[SM]" (HMIHY) is a natural language human-computer spoken dialogue system developed at AT&T Research Labs. The system enables AT&T customers to interact verbally with an automated agent over the phone. Callers can ask for their account balance, help with AT&T rates and calling plans, explanations of certain bill charges, or identification of numbers. Speech data collected from the deployed system has been assembled into a corpus of human-computer dialogues. The HMIHY corpus contains 5,690 complete human-computer dialogues that collectively contain 20,013 caller turns. Each caller turn in the corpus was annotated with one of seven emotional labels: *positive/neutral, somewhat frustrated, very frustrated, somewhat angry, very angry, somewhat other negative*[2]*, very other negative*. However, the distribution of the labels was so skewed (73.1% were labeled as *positive/neutral*) that the emotions were collapsed to *negative* and *non-negative*.

In addition to the set of automatic acoustic-prosodic features found to be useful for emotional classification of the EPSaT corpus, the features I examined in the HMIHY corpus were designed to exploit the discourse information available in the domain of spontaneous human-machine conversation. Transcriptive features – lexical items, filled pauses, and non-speech human noises – we recorded as features, as too were the dialogue acts of each caller turn. In addition, I included contextual features that were designed to track the history of the previously mentioned features over the course of the dialogue. Specifically, contextual information included the rate of change of the acoustic-prosodic features of the previous two turns plus the transcriptive and pragmatic features of the previous two turns as well.

The corpus was divided into training (75%) and testing (25%) sets. The machine learning algorithm employed was BOOSTEXTER, an algorithm that forms a hypothesis by combining the results of several iterations of weak-learner decisions. Classification accuracy using the automatic acoustic-prosodic features was recorded to be approximately 75%. The majority class baseline (always guessing *non-negative*) was 73%. By adding the other feature-sets one by one, prediction accuracy was iteratively improved, as described more fully in (Liscombe et al., 2005b). Using all the features combined – acoustic-prosodic, lexical, pragmatic, and contextual – the resulting classification accuracy was 79%, a healthy 8% improvement over baseline performance and a 5% improvement over the automatic acoustic-prosodic features alone.

2.3 ITSpoke

This section describes more recent research I have been conducting with the University of Pittsburgh's Intelligent Tutoring Spoken Dialogue System (ITSpoke) (Litman and Silliman, 2004). The goal of this research is to wed spoken language technology with instructional technology in order to promote learning gains by enhancing communication richness. ITSpoke is built upon the Why2-Atlas tutoring back-end (VanLehn et al., 2002), a text-based Intelligent Tutoring System designed to tutor students in the domain of qualitative physics using natural language interaction. Several corpora have been recorded for development of ITSpoke, though most of the work presented here involves tutorial data between a student and human tutor. To date, we have labeled the human-human corpus for anger, frustration, and uncertainty.

As this work is an extension of previous work, I chose to extract most of the same features I had extracted from the EPSaT and HMIHY corpora. Specifically, I extracted the same set of automatic acoustic-prosodic features, as well as contextual features measuring the rate of change of acoustic-prosodic features of past student turns. A new feature set was introduced as well, which I refer to as the breath-group feature set, and which is an automatic method for segmenting utterances into intonationally meaningful units by identifying pauses using background noise estimation. The breath group feature set

[2] 'Other negative' refers to any emotion that is perceived negatively but is not anger nor frustration.

comprises the number of breath-groups in each turn, the pause time, and global acoustic-prosodic features calculated for the first, last, and longest breath-group in each student turn.

I used the WEKA machine learning software package to classify whether a student answer was perceived to be *uncertain*, *certain*, or *neutral*[3] in the ITSpoke human-human corpus. As a predictor, C4.5, a decision-tree learner, was boosted with AdaBoost, a learning strategy similar to the one presented in Section 2.2. The data were randomly split into a training set (90%) and a testing set (10%). The automatic acoustic-prosodic features performed at 75% accuracy, a relative improvement of 13% over the baseline performance of always guessing *neutral*. By adding additional feature-sets – contextual and breath-group information – I observed an improved prediction accuracy of 77%. Thus indicating that breath-group features are useful. I refer the reader to (Liscombe et al., 2005a) for in-depth implications and further analysis of these results. In the immediate future, I will extract features previously mentioned in Section 2.2 as well as the exploratory features I will discuss in the following section.

3 Work-in-progress

In this section I describe research I have begun to conduct and plan to complete in the coming year, as agreed-upon in February, 2006 by my dissertation committee. I will explore features that are not well studied in emotion classification research, primarily pitch contour and voice quality approximation. Furthermore, I will outline how I plan to implement and evaluate an emotion detection and response module into ITSpoke.

3.1 Pitch Contour Clustering

The global acoustic-prosodic features used in most emotion prediction studies capture meaningful prosodic variation, but are not capable of describing the linguistically meaningful intonational behavior of an utterance. Though phonological labeling methods exist, such as ToBI, annotation of this sort is time-consuming and must be done manually. Instead, I propose an automatic algorithm that directly compares pitch contours and then groups them into classes based on abstract form. Specifically, I intend to use partition clustering to define a disjoint set of similar prosodic contour types over our data. I hypothesize that the resultant clusters will be theoretically meaningful and useful for emotion modeling. The similarity metric used to compare two contours will be edit distance, calculated using dynamic time warping techniques. Essentially, the algorithm finds the best fit between two contours by stretching and shrinking each contour as necessary. The score of a comparison is calculated as the sum of the normalized real-valued distances between mapped points in the contours.

[3]With respect to certainness.

3.2 Voice Quality

Voice quality is a term used to describe a perceptual coloring of the acoustic speech signal and is generally believed to play an important role in the vocal communication of emotion. However, it has rarely been used in automatic classification experiments because the exact parameters defining each quality of voice (*e.g.*, creaky and breathy) are still largely unknown. Yet, some researchers believe much of what constitutes voice quality can be described using information about glottis excitation produced by the vocal folds, most commonly referred to as the glottal pulse waveform. While there are ways of directly measuring the glottal pulse waveform, such as with an electroglottograph, these techniques are too invasive for practical purposes. Therefore, the glottal pulse waveform is usually approximated by inverse filtering of the speech signal. I will derive glottal pulse waveforms from the data using an algorithm that automatically identifies voiced regions of speech, obtains an estimate of the glottal flow derivative, and then represents this using the Liljencrants-Fant parametric model. The final result is a glottal pulse waveform, from which features can be extracted that describe the shape of this waveform, such as the Open and Skewing Quotients.

3.3 Implementation

The motivating force behind much of the research I have presented herein is the common assumption in the research community that emotion modeling will improve spoken dialogue systems. However, there is little to no empirical proof testing this claim (See (Pon-Barry et al., In publication) for a notable exception.). For this reason, I will implement functionality for detecting and responding to student emotion in ITSpoke (the Intelligent Tutoring System described in Section 2.3) and analyze the effect it has on student behavior, hopefully showing (quantitatively) that doing so improves the system's effectiveness.

Research has shown that frustrated students learn less than non-frustrated students (Lewis and Williams, 1989) and that human tutors respond differently in the face of student uncertainty than they do when presented with certainty (Forbes-Riley and Litman, 2005). These findings indicate that emotion plays an important role in Intelligent Tutoring Systems. Though I do not have the ability to alter the discourse-flow of ITSpoke, I will insert active listening prompts on the part of ITSpoke when the system has detected either frustration or uncertainty. Active listening is a technique that has been shown to diffuse negative emotion in general (Klein et al., 2002). I hy-

pothesize that diffusing user frustration and uncertainty will improve ITSpoke.

After collecting data from an emotion-enabled ITSpoke I will compare evaluation metrics with those of a control study conducted with the original ITSpoke system. One such metric will be learning gain, the difference between student pre- and post-test scores and the standard metric for quantifying the effectiveness of educational devices. Since learning gain is a crude measure of academic achievement and may overlook behavioral and cognitive improvements, I will explore other metrics as well, such as: the amount of time taken for the student to produce a correct answer, the amount of negative emotional states expressed, the quality and correctness of answers, the willingness to continue, and subjective post-tutoring assessments.

4 Contributions

I see the contributions of my dissertation to be the extent to which I have helped to answer the questions I posed at the outset of this paper.

4.1 How is emotion communicated in speech?

The experimental design of extracting features from spoken utterances and conducting machine learning experiments to predict emotion classes identifies features important for the vocal communication of emotion. Most of the features I have described here are well established in the research community; statistic measurements of fundamental frequency and energy, for example. However, I have also described more experimental features as a way of improving upon the state-of-the-art in emotion modeling. These exploratory features include breath-group segmentation, contextual information, pitch contour clustering, and voice quality estimation. In addition, exploring three domains will allow me to comparatively analyze the results, with the ultimate goal of identifying universal qualities of spoken emotions as well as those that may particular to specific domains. The findings of such a comparative analysis will be of practical benefit to future system builders and to those attempting to define a universal model of human emotion alike.

4.2 Does emotion modeling help?

By collecting data of students interacting with an emotion-enabled ITSpoke, I will be able to report quantitatively the results of emotion modeling in a spoken dialogue system. Though this is the central motivation for most researchers in this field, there is currently no definitive evidence either supporting or refuting this claim.

References

M. E. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, 2005. *Prosodic Typology – The Phonology of Intonation and Phrasing*, chapter 2 The original ToBI system and the evolution of the ToBI framework. Oxford, OUP.

J. R. Davitz, 1964. *The Communication of Emotional Meaning*, chapter 8 Auditory Correlates of Vocal Expression of Emotional Feeling, pages 101–112. New York: McGraw-Hill.

Kate Forbes-Riley and Diane J. Litman. 2005. Using bigrams to identify relationships between student certainness states and tutor responses in a spoken dialogue corpus. In *Proceedings of 6th SIGdial Workshop on Discourse and Dialogue,*, Lisbon, Portugal.

J. Klein, Y. Moon, and R. W. Picard. 2002. This computer responds to user frustration: Theory, design, and results. *Interacting with Computers*, 14(2):119–140, February.

V. E. Lewis and R. N. Williams. 1989. Mood-congruent vs. mood-state-dependent learning: Implications for a view of emotion. *D. Kuiken (Ed.), Mood and Memory: Theory, Research, and Applications, Special Issue of the Journal of Social Behavior and Personality*, 4(2):157–171.

Jackson Liscombe, Jennifer Venditti, and Julia Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Eurospeech*, Geneva, Switzerland.

Jackson Liscombe, Julia Hirschberg, and Jennifer Venditti. 2005a. Detecting certainness in spoken tutorial dialogues. In *Proceedings of Interspeech*, Lisbon, Portugal.

Jackson Liscombe, Guiseppe Riccardi, and Dilek Hakkani-Tür. 2005b. Using context to improve emotion detection in spoken dialogue systems. In *Proceedings of Interspeech*, Lisbon, Portugal.

Diane Litman and Scott Silliman. 2004. Itspoke: An intelligent tutoring spoken dialogue system. In *Proceedings of the 4th Meeting of HLT/NAACL* (Companion Proceedings), Boston, MA, May.

Heather Pon-Barry, Karl Schultz, Elizabeth Owen Bratt, Brady Clark, and Stanley Peters. In publication. Responding to student uncertainty in spoken tutorial dialogue systems. *International Journal of Artificial Intelligence in Education (IJAIED)*.

K. R. Scherer, J. Koivumaki, and R. Rosenthal. 1972. Minimal cues in the vocal communication of affect: Judging emotions from content-masked speech. *Journal of Psycholinguistic Research*, 1:269–285.

K. VanLehn, P. Jordan, and C. P. Rose. 2002. The architecture of why2-atlas: A coach for qualitative physics essay writing. In *Proceedings of the Intelligent Tutoring Systems Conference*, Biarritz, France.

Document Representation and Multilevel Measures of Document Similarity

Irina Matveeva
Dept. of Computer Science
University of Chicago
matveeva@cs.uchicago.edu

Abstract

We present our work on combining large-scale statistical approaches with local linguistic analysis and graph-based machine learning techniques to compute a combined measure of semantic similarity between terms and documents for application in information extraction, question answering, and summarisation.

1 Introduction

Document indexing and representation of term-document relations are crucial for document classification, clustering and retrieval. In the traditional bag-of-words vector space representation of documents (Salton and McGill, 1983) words represent orthogonal dimensions which makes an unrealistic assumption about their independence.

Since document vectors are constructed in a very high dimensional vocabulary space, there has been a considerable interest in low-dimensional document representations to overcome the drawbacks of the bag-of-words document vectors. Latent Semantic Analysis (LSA) (Deerwester et al., 1990) is one of the best known dimensionality reduction algorithms in information retrieval.

In my research, I consider different notions of similarity measure between documents. I use dimensionality reduction and statistical co-occurrence information to define representations that support them.

2 Dimensionality Reduction for Document and Term Representation

A vector space representation of documents is very convenient because it puts documents in a Euclidean space where similarity measures such as inner product and cosine similarity or distance are immediately available. However, these measures will not be effective if they do not have a natural interpretation for the original text data.

I have considered several approaches to computing a vector space representation of text data for which inner product and distance make sense. The general framework is to construct a matrix of pairwise similarities between terms or documents and use appropriate methods of dimensionality reduction to compute low dimensional vectors. The inner product between the resulting vectors must preserve the similarities in the input matrix. The similarities matrix can be computed using different notions of similarity in the input space. Different dimensionality reduction techniques impose different conditions on how the similarities are preserved.

I investigated how external query-based similarity information can be used to compute low dimensional document vectors. Similar to LSA, this approach used weighted bag-of-words document vectors as input which limited its effectiveness. The next step was to develop the Generalized Latent Semantic Analysis framework that allows to compute semantically motivated term and document vectors.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 235–238,
New York, June 2006.

2.1 Document Representation with the Locality Preserving Projection Algorithm

The Locality Preserving Projection algorithm (LPP) (He and Niyogi, 2003) is a graph-based dimensionality reduction algorithm that computes low dimensional document vectors by preserving local similarities between the documents. It requires a vector space representation of documents as input. In addition, it uses the adjacency matrix of the nearest neighbors graph of the data. It can be shown, see (He and Niyogi, 2003), that the Euclidean distance in the LPP space corresponds to similarity in the document space.

The information about the similarity of the input documents is contained in the adjacency matrix of the nearest neighbors graph. In this graph, nodes represent documents and are connected by an edge if the documents are similar. This graph can be constructed using *any* similarity measure between the documents, for example, the query-based similarity between the documents obtained from relevance feedback. The base case is to use inner products between the input document vectors and to connect k nearest neighbors.

We considered several ways of modifying the graph, see (Matveeva, 2004). We used relevance feedback and pseudo relevance feedback from the base line term matching retrieval to identify the top N documents most related to the query. We added edges to the document neighborhood graph to connect these N documents. Our experiments showed that incorporating this external relevance information into the LPP graph improves the performance on the information retrieval tasks, in particular at high levels of recall. Without the use of external information, the performance of the LPP algorithm was comparable to the performance of the LSA algorithm up to recall of 0.6–0.7. At higher levels of recall, LSA achieves a precision that is about 0.1 better than LPP. The precision at high levels of recall seemed to be a weak point of LPP. Fortunately, using the relevance feedback helped to improve the performance in particular in this range of recall.

We found the LPP algorithm to be very sensitive to the graph structure. It confirmed the intuition that the Euclidean distance between the document vectors in the bag-of-words representation is not a good similarity measure. When we added query relevance information to the graph, we introduced a similarity metric on the document space that was closer to the true similarity. However, this information was only partial, because only a subset of the edges reflected this true similarity. The next step was therefore to develop a vector space representation for documents which did not require the bag-of-words representation as input.

2.2 Generalized Latent Semantic Analysis

We developed the Generalized Latent Semantic Analysis (GLSA) framework to compute semantically motivated term and document vectors (Matveeva et al., 2005). We begin with semantically motivated pair-wise term similarities and use dimensionality reduction to compute a vector space representation for terms. Our approach is to focus on similarity between vocabulary terms. We compute representations and similarities for terms and consider documents to be linear combinations of terms. This shift from dual document-term representation to terms has the following motivation.

- Terms offer a much greater flexibility in exploring similarity relations than documents. The availability of large document collections such as the Web offers a great resource for statistical approaches. Recently, co-occurrence based measures of semantic similarity between terms has been shown to improve performance on such tasks as the synonymy test, taxonomy induction, etc. (Turney, 2001; Terra and Clarke, 2003; Chklovski and Pantel, 2004). On the other hand, many semi-supervised and transductive methods based on document vectors cannot yet handle such large document collections.

- While the vocabulary size is still quite large, it is intuitively clear that the intrinsic dimensionality of the vocabulary space is much lower. Content bearing words are often combined into semantic classes that correspond to particular activities or relations and contain synonyms and semantically related words. Therefore, it seems very natural to represent terms as low dimensional vectors in the space of semantic concepts.

2.2.1 GLSA Algorithm

The GLSA algorithm takes as input a document collection C with vocabulary V and a large corpus W. It has the following outline:

1. Construct the weighted term document matrix D based on C
2. For the vocabulary words in V, obtain a matrix of pair-wise similarities, S, using the large corpus W
3. Obtain the matrix U^T of low dimensional vector space representation of terms that preserves the similarities in S, $U^T \in R^{k \times |V|}$. The columns of U^T are k-dimensional term vectors
4. Compute document vectors by taking linear combinations of term vectors $\hat{D} = U^T D$

In step 2 of the GLSA algorithm we used pointwise mutual information (PMI) as the co-occurrence based measure of semantic associations between pairs of the vocabulary terms. We used the singular value decomposition in step 3 to compute GLSA term vectors.

2.2.2 Experimental Evaluation

We used the TOEFL, TS1 and TS2 synonymy tests to demonstrate that the GLSA vector space representation for terms captures their semantic relations, see (Matveeva et al., 2005) for details. Our results demonstrate that similarities between GLSA term vectors achieve better results than PMI scores and outperform the related PMI-IR approach (Turney, 2001; Terra and Clarke, 2003). On the TOEFL test GLSA achieves the best precision of 0.86, which is much better than our PMI baseline as well as the highest precision of 0.81 reported in (Terra and Clarke, 2003). GLSA achieves the same maximum precision as in (Terra and Clarke, 2003) for TS1 (0.73) and higher precision on TS2 (0.82 compared to 0.75 in (Terra and Clarke, 2003)).

We also conducted document classification experiments to demonstrate the advantage of the GLSA document vectors (Matveeva et al., 2005). We used a k-nearest neighbors classifier for a set of 5300 documents from 6 dissimilar groups from the 20 news groups data set. The k-nn classifier achieved higher accuracy with the GLSA document vectors than with the traditional tf-idf document vectors, especially with fewer training examples. With 100 training examples, the k-nn classifier with GLSA had 0.75 accuracy vs. 0.58 with the tf-idf document vectors. With 1000 training examples the numbers were 0.81 vs. 0.75.

The inner product between the GLSA document vectors can be used as input to other algorithms. The language modelling approach (Berger and Lafferty, 1999) proved very effective for the information retrieval task. Berger et. al (Berger and Lafferty, 1999) used translation probabilities between the document and query terms to account for synonymy and polysemy. We proposed to use low dimensional term vectors for inducing the translation probabilities between terms (Matveeva and Levow, 2006). We used the same k-nn classification task as above. With 100 training examples, the k-nn accuracy based on tf-idf document vectors was 0.58 and with the similarity based on the language modelling with GLSA term translation probabilities the accuracy was 0.69. With larger training sets the difference in performance was less significant. These results illustrate that the pair-wise similarities between the GLSA term vectors add important semantic information which helps to go beyond term matching and deal with synonymy and polysemy.

3 Work in Progress

Many recent applications such as document summarization, information extraction and question answering require a detailed analysis of semantic relations between terms within and across documents and sentences. Often one has a number of sentences or paragraphs and has to choose the candidate with the highest level of relevance for the topic or question. An additional requirement may be that the information content of the next candidate is different from the sentences that are already chosen.

In these cases, it seems natural to have different levels of document similarity. Two sentences or paragraphs can be similar because they contain information about the same people or events. In this case, the similarity can be based on the number of the named entities they have in common. On the other hand, they can be similar because they contain synonyms or semantically related terms.

I am currently working on a combination of similarity measures between terms to model document similarity. I divide the vocabulary into general vocabulary terms and named entities and compute a separate similarity score for each group of terms. The overall document similarity score is a function of these two scores. To keep the vocabulary size manageable and denoise the data, we only use the content bearing words from the set of the general vocabulary terms. We use a parser to identify nouns and adjectives that participate in three types of syntactic relations: subject, direct object, the head of the noun phrase with an adjective or noun as a modifier for nouns and the modifier of a noun for adjectives. Currently we include only such nouns and adjectives in the set of the content bearing vocabulary terms.

We used the TDT2 collection for preliminary classification experiments. We used a k-nn classifier to classify documents from the 10 most frequent topics. We used tf-idf document vectors indexed with 55,729 general vocabulary words as our baseline. The set of the content bearing words was much smaller and had 13,818 nouns and adjectives. The GLSA document vectors improved the classification accuracy over the baseline and outperformed LSA document vectors. This validates our approach to selecting the content bearing terms and shows the advantage of using the GLSA framework. We are going to extend the set of content bearing words and to include verbs. We will take advantage of the flexibility provided by our framework and use syntax based measure of similarity in the computation of the verb vectors, following (Lin, 1998).

Currently we are using string matching to compute the named entity based measure of similarity. We are planning to integrate more sophisticated techniques in our framework.

4 Conclusion

We developed the GLSA framework for computing semantically motivated term and document vectors. This framework takes advantage of the availability of large document collections and recent research of corpus-based term similarity measures and combines them with dimensionality reduction algorithms.

Different measures of similarity may be required for different groups of terms such as content bearing vocabulary words and named entities. To extend the GLSA approach to computing the document vectors, we use a combination of similarity measures between terms to model the document similarity. This approach defines a fine-grained similarity measure between documents and sentences. Our goal is to develop a multilevel measure of document similarity that will be helpful for summarization and information extraction.

References

Adam Berger and John Lafferty. 1999. Information retrieval as statistical translation. In *Proc. of the 22rd ACM SIGIR*.

Timothy Chklovski and Patrick Pantel. 2004. Verbocean: Mining the web for fine-grained semantic verb relations. In *Proc. of EMNLP*.

Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407.

Xiaofei He and Partha Niyogi. 2003. Locality preserving projections. In *Proc. of NIPS*.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *COLING-ACL*, pages 768–774.

Irina Matveeva and Gina-Anne Levow. 2006. Computing term translation probabilities with generalized latent semantic analysis. In *Proc. of EACL*.

Irina Matveeva, Gina-Anne Levow, Ayman Farahat, and Christian Royer. 2005. Generalized latent semantic analysis for term representation. In *Proc. of RANLP*.

Irina Matveeva. 2004. Text representation with the locality preserving projection algorithm for information retrieval task. In *Master's Thesis*.

Gerard Salton and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. McGraw-Hill.

Egidio L. Terra and Charles L. A. Clarke. 2003. Frequency estimates for statistical word similarity measures. In *Proc.of HLT-NAACL*.

Peter D. Turney. 2001. Mining the web for synonyms: PMI–IR versus LSA on TOEFL. *Lecture Notes in Computer Science*, 2167:491–502.

Logical investigations on the adequacy of certain feature-based theories of natural language

Anders Søgaard
Center for Language Technology
Njalsgade 80
DK-2300 Copenhagen
anders@cst.dk

Abstract

A theory of natural language can be evaluated on both extensional and intensional grounds. Systematic investigations of the extension of a theory may, for instance, lead to studies of the invariance properties of such theories. The intentional parameters that I wish to address include complexity, learnability, and monotonicity. The main results, on which my thesis builds, up to this point, include: (i) the universal recognition problem of model-theoretic feature-based grammar formalisms is complete for non-deterministic polynomial time, since such formalisms have the polysize model property, (ii) this result holds also for linearization-based extensions, (iii) the universal recognition problem of strongly monotonic, hybrid feature-based grammar formalisms is decidable in deterministic polynomial time, and (iv) there exists a strongly monotonic unification categorial grammar that is learnable in the limit from positive data. In addition, invariance studies have lead to the identification of a class of modal languages that define common feature-based grammar formalisms. The objective of my studies is to identify a tractable and learnable feature-based formalism.

1 Introduction

My work addresses certain extensional and intensional properties of various feature-based theories of natural language, incl. unification categorial grammar (Zeevat, 1988) and head-driven phrase structure grammar (Pollard and Sag, 1994). The theories are referred to henceforth as UCG and HPSG. A feature-based theory of natural language defines a set of feature-based grammars (and interfaces). A feature-based grammar associates feature structures with the strings of the language in question. Consequently, it makes sense to start off with a definition of a feature structure. A signature $\langle \mathsf{Lbls}, \mathsf{Atmc} \rangle$ is a pair of sets of labels and atomic informations. In UCG and HPSG, both are finite. A feature structure of a signature $\langle \mathsf{Lbls}, \mathsf{Atmc} \rangle$ is then an ordered triple $\langle \mathbb{N}, \{R_\lambda\}_{\lambda \in \mathsf{Lbls}}, \{Q_\alpha\}_{\alpha \in \mathsf{Atmc}} \rangle$, where $\mathbb{N}$ is a set of nodes, R_λ is a partial function, and Q_α is a unary one, for all $\lambda \in \mathsf{Lbls}$ and $\alpha \in \mathsf{Atmc}$.

Grammars employ feature structures in different ways. A *hybrid* grammar, in its most raw format, is a 4-tuple $\langle \langle \mathsf{Lbls}, \mathsf{Atmc} \rangle, \mathbb{V}, \mathsf{Rules}, \mathtt{start} \rangle$, where $\mathbb{V}$ is the vocabulary. One may add a specification function such that, for instance, $\forall x \; \exists y \; R_\lambda(x, y) \rightarrow Q_\alpha(y)$. Intuitively, if $\mathcal{L}(\mathcal{G})$ is the language of $\mathcal{G}$, and if $\mathcal{G}$ is a hybrid grammar, $\mathcal{L}(\mathcal{G})$ is the set of strings "modelled" (derivable) by the grammar, whereas the language of a model-theoretic grammar is the set of strings that (or whose relational structures) model the grammar. The generative-enumerative core of a hybrid grammar is in its set of rules (Rules).

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 239–242,
New York, June 2006.

The language of $\mathcal{G}$ is defined as

$$\mathcal{L}(\mathcal{G}) = \{x \in \mathbb{V}^* | \exists f \in \mathcal{F}\ \texttt{start} \sqsubseteq f \wedge f \Rightarrow x\}$$

where $\mathcal{F}$ is the set of feature structures, and $f' \sqsubseteq f$ means that the information in f' is also in f, and $f \wedge f'$ is consistent. Finally, $f \Rightarrow \sigma$ means that σ is derivable from f by Rules. It should be obvious that a hybrid grammar is "hybrid" in the sense that it combines generative rules and subsumption ($\sqsubseteq$), which is essentially model-theoretically defined, i.e. $f' \sqsubseteq f$ iff if $f \models \phi$ then $f' \models \phi$.

Consider a grammar $\mathcal{G}'$, which is entirely model-theoretic, i.e. the grammar is axiomatically defined, and feature structures are seen as Kripke frames. In other words, a *model-theoretic* grammar is a 4-tuple $\langle\langle \mathsf{Lbls}, \mathsf{Atmc}\rangle, \mathbb{V}, \mathsf{Axms}, \texttt{root}\rangle$, where rules have been replaced with a set of axioms Axms defined in some logic, and the `start` category is replaced with a `root` proposition, defined in the axioms. The signature is now a signature of modalities and propositions. It is important, in order to maintain the overall picture, to remember that on the standard translation into first order logic, modalities and propositions translate into binary and unary relations, respectively. Consequently, this is, at the moment, just a notational change. The introduction of modal vocabulary is relevant to the specification of feature-based theories later on.

The universal recognition problem amounts to this question: Given some pair $\sigma, \mathcal{G}$, $\sigma \in \mathcal{L}(\mathcal{G})$? In particular, when it is said that the universal recognition problem of some formalism is in some complexity class, it means that there exists an algorithm such that the membership of any string in any grammar licensed by the formalism can be decided in the time complexity of that class by running the algorithm. The universal recognition problems of model-theoretic UCG and HPSG, and the linearization-based extension of the latter, and strongly monotonic HPSG are examined in a minute.

Our introductions of UCG and HPSG are of course only partial, since this paper is of limited length. In fact, no more than a paragraph is spend on these introductions:

Unification categorial grammar UCG and HPSG are both said to be sign-based, i.e. the fundamental unit is the sign. A sign in UCG has the structure W:C:S:O, where W contains information about the phonology of the sign, C presents its syntactic category, S is the semantics, and O constrains word order in determining how the sign combines with other signs. Signs combine by functional application (instantiation and stripping). Instantiation checks if the active part of the syntactic category of the functor unifies with the syntactic category of the argument, and if unification succeeds, the instantiated functor is stripped, and the phonology features are concatenated. Type hierarchies extend UCG in a natural way. Instantiation and stripping can be interpreted as phrasal types rather than functions. Model-theoretic parsing of some string $\sigma \in \mathcal{L}(\mathcal{G})$ then amounts to finding a connected and rooted (minimal) model $\mathcal{M}$ whose linearization is σ, s.t. $\mathcal{M}, w \in [\![\texttt{root}]\!] \models \texttt{Axms}$.

Head-driven phrase structure grammar HPSG parsing is much the same, except Axms is conjoined with Prncp, the set of linguistic principles. One traditional problem with HPSG is that it employs sets. Some recent (computationally oriented) versions of HPSG substitute sets with so-called "diff-lists", which are briefly lists with pointers to their last elements, and for now we settle with this option. An alternative is mentioned in our discussion of linearization-based HPSG, namely a simulation of sets as underspecified lists; or one can perhaps employ polyadic modalities (n-ary relations). The linguistic principles in Prncp include, for instance, the head feature principle, which says that in a headed phrase, the HEAD value of the mother is identical to that of the head daughter, the immediate dominance principle and the weak coordination principle.

2 Some formal results

Our first complexity result, i.e. (i) the universal recognition problem of model-theoretic feature-based grammar formalisms is complete for non-deterministic polynomial time, since such formalisms have the polysize model property, is

obtained by specification of UCG and HPSG in some modal language that has a model checking problem of polynomial time complexity. The model checking amounts to evaluating a formula ϕ in a model $\mathcal{M}$. If a formalism has the polysize model property, its universal recognition problem can be evaluated on small models that are polynomial in the size of the strings. If the specification language has a polynomial model checking problem, a model can thus be non-deterministically chosen and evaluated in polynomial time, and the result follows. Consequently, the quest is two-fold: It is necessary to establish the polysize model property for UCG and HSPG, and we then need to identify an adequate specification language that embeds these theories. The polysize model property follows from Lemma 2.1.[1]

Lemma 2.1. *Say ϕ represents a* UCG *or* HPSG *recognition problem for a string σ. If there exists a model $\mathcal{M}$ and a node $w \in \mathbb{N}$ s.t. $\mathcal{M}, w \in [\![\texttt{root}]\!] \models \phi$, then there also exists a model $\mathcal{M}'$ of at most k cardinality and a node $w' \in \mathbb{N}$ s.t. $\mathcal{M}', w' \in [\![\texttt{root}]\!] \models \phi$, where $k = (2|\sigma| - 1) \times (u + 1) \times \mathbf{paths}$, where u is the number of unary rules in the grammar, and* **paths** *is a constant that depends on the non-recursive part of the feature geometries of* UCG *and* HPSG. *In particular,* $\mathbf{paths} = |\{\pi \in \texttt{Lbls}^* | \textit{no label occurs twice in } \pi\}|$.

It is now left to show that UCG and HPSG can be specified (defined) in some formal language that has a polynomial time model checking problem. Since UCG subsumes HPSG, it suffices to show that this holds for HPSG. Various translations of HPSG into specification languages have been proposed, and my recent work includes a couple of such translations, but in this synopsis, to save space, we refer to the specification language of Kracht (1995). He defines a translation of HPSG into $\mathsf{PDL}^{\cup,[*]}$, propositional dynamic logic with intersection and the master modality. The master modality is defined s.t. $\mathcal{M}, w \models [*]\phi$ iff $\forall w'((w, w') \in (\bigcup_{\alpha \text{ is atomic}} R_\alpha) \& \mathcal{M}, w' \models \phi)$. It is trivial to show the high undecidability of this language, for instance, the recurrent tiling problem can be encoded in $\mathsf{PDL}^{\cup,[*]}$. The model checking of $\mathsf{PDL}^{\cup,[*]}$ is indeed decidable in polynomial time; this is evident from the investigations of Lange (2006). Consequently, Theorem 2.2 follows.

Theorem 2.2. *The universal recognition problem of* UCG *and* HPSG *is decidable in non-deterministic polynomial time.*

The result can be extended to linearization-based versions of model-theoretic UCG and HPSG. On the model-theoretic perspective, immediate dominance and linear precedence are already split, since they are represented by different modalities. The thing to do when languages of freer word-order are considered, is then simply to relax the linearization of immediate dominance principles. The master modality of $\mathsf{PDL}^{\cup,[*]}$ can be used to implement weak linear precedence. Weak linear precedence is thus, in some sense, constraints on an underspecified list of strings, and domain union, for instance, is "unification" of underspecified lists.

Strong monotonicity has been mentioned a couple of times. The notion is relevant on a hybrid set-up. Some grammar formalisms are non-monotonic in the traditional sense, but we confine ourselves to monotonic ones, for the simple reason that the modal languages considered here are all monotonic. The notion of *strong* monotonicity is different. Consider a conventional context-free grammar. On our definition of strong monotonicity, a context-free grammar $\mathcal{G}$ of $\mathcal{L}$ is *not* strongly monotonic if it is ambiguous on $\mathcal{L}$, i.e. if there exists a string $\sigma \in \mathcal{L}$, such that more than one tree can be derived by $\mathsf{Rules}_\mathcal{G}$. The strong monotonicity hypothesis, i.e. that natural language grammars are strongly monotonic, is very strong and somewhat unnatural to most linguists. Since Linguistics 101, we were taught that languages are inherently ambiguous. Strongly monotonic grammars of course have formal interest, since they exhibit a number of nice properties, discussed in the next paragraph, but they *need* not be irrelevant

[1] Unary rules only apply once to the same unary extension in Lemma 2.1. In the proof of Theorem 2.3, a unary extension is the result of a single application, i.e. $u = 1$ in Lemma 2.1. It is not clear to me what the linguistic relevant restriction is.

in linguistics either. In feature-based grammars that employ type hierarchies, it is possible, after all, to underspecify ambiguities. It has been argued that such underspecification is possible and a linguistically interesting option in the context of both quantification, attachment ambiguities, and the combinatorics of case and word order.

Say $\mathcal{G}$ is a hybrid grammar and strongly monotonic. For one thing, this means that the lexicon in $\mathsf{Rules}_{\mathcal{G}}$ is rigid s.t. a partial function map strings onto feature structures. It also means that ϕ has a *unique* model of size less than or equal to k. A rather restrictive parsing algorithm is introduced: Say $\mathsf{Rules}_{\mathcal{G}}$ consists of b binary rules and u unary ones. $\mathcal{G}$ tries to combine pairs of constituents bottom-up by b, and if this does not succeed, u is used to extend any of the constituents by a single application. On the assumption that Rules contains no unary rules, $(\frac{|\sigma|^2-|\sigma|}{2}) \times b$ is the number of possible projections. When unary rules are added, this number is multiplied by the number of unary rules times the number of binary rules, since the binary rules are first tried out, and if that doesn't work, unary rules are used to extend nodes, and binary rules are applied again. The algorithm only has to run once because of strong monotonicity. Consequently, Theorem 2.3 holds:

Theorem 2.3. *The universal recognition problem of strongly monotonic and hybrid feature-based grammars is decidable in deterministic polynomial time.*

Proof. $(\frac{|\sigma|^2-|\sigma|}{2}) \times b$ is the number of possible projections in the abscence of unary rules. Add unary rules and the number of possible projections is

$$\frac{b\times(|\sigma|^2-|\sigma|)}{2}(u+1)^3 + |\sigma|u$$

For each step, unification is tested. Unification is decidable in time $\Theta(\delta \times \omega(\delta))$ (Hegner, 1991), where δ is the number of distinct edges in the two feature structures, i.e. $\delta = \mathbf{paths}$ in the above, and $\omega(\delta)$ is the inverse Ackermann function. For all practical purposes, $\omega(\delta)$ is lower than 5 (Hegner, 1991). Nothing else has to be computed to decide universal recognition for a strongly monotonic hybrid feature-based grammar. The result follows. □

The learnability result, "(iv)" in the above, derives from a result established by Kanazawa (1998), namely that rigid categorial grammars are leanable in the limit, even from positive data. If so it follows that there exists strongly monotonic unification categorial grammars that are also learnable in the limit from positive data, since strongly monotonic grammars are rigid, by definition, and since simple unification categorial grammars can be embedded in classical ones.

I envisage a tractable and learnable feature-based grammar formalism to look much like strongly monotonic UCG extended with type hiearchies and linearization. The notion of strong monotonic can be relativized in various ways without loosing tractability, and this line of research should be pursued.

References

Stephen Hegner. 1991. Horn-extended feature structures. In *The 5th European Chapter of the Association for Computational Linguistics*, pages 33–38, Berlin, Germany.

Makoto Kanazawa. 1998. *Learnable classes of categorial grammars*. CSLI Publications, Stanford, California.

Marcus Kracht. 1995. Is there a genuine modal perspective on feature structures? *Linguistics & Philosophy*, 18:401–458.

Martin Lange. 2006. Model checking propositional dynamic logic with all extras. *Journal of Applied Logic*, 4:39–49.

Carl Pollard and Ivan Sag. 1994. *Head-driven phrase structure grammar*, volume 4 of *Studies in Contemporary Linguistics*. The University of Chicago Press, Chicago, Illinois.

Henk Zeevat. 1988. Combining categorial grammar and unification. In Uwe Reyle and Christian Rohrer, editors, *Natural language parsing and linguistic theories*, pages 202–229. Reidel, Dordrecht, Germany.

A Hybrid Approach to Biomedical Named Entity Recognition and Semantic Role Labeling

Richard Tzong-Han Tsai

Department of Computer Science and Information Engineering

National Taiwan University

Nankang, Taipei, Taiwan, 115

thtsai@iis.sinica.edu.tw

Abstract

In this paper, we describe our hybrid approach to two key NLP technologies: biomedical named entity recognition (Bio-NER) and (Bio-SRL). In Bio-NER, our system successfully integrates linguistic features into the CRF framework. In addition, we employ web lexicons and template-based post-processing to further boost its performance. Through these broad linguistic features and the nature of CRF, our system outperforms state-of-the-art machine-learning-based systems, especially in the recognition of protein names (F=78.5%). In Bio-SRL, first, we construct a proposition bank on top of the popular biomedical GENIA treebank following the PropBank annotation scheme. We only annotate the predicate-argument structures (PAS's) of thirty frequently used biomedical verbs (predicates) and their corresponding arguments. Second, we use our proposition bank to train a biomedical SRL system, which uses a maximum entropy (ME) machine-learning model. Thirdly, we automatically generate argument-type templates, which can be used to improve classification of biomedical argument roles. Our experimental results show that a newswire English SRL system that achieves an F-score of 86.29% in the newswire English domain can maintain an F-score of 64.64% when ported to the biomedical domain. By using our annotated biomedical corpus, we can increase that F-score by 22.9%. Adding automatically generated template features further increases overall F-score by 0.47% and adjunct (AM) F-score by 1.57%, respectively.

1 Introduction

The volume of biomedical literature available on the Web has experienced unprecedented growth in recent years, and demand for efficient methods to process this material has increased accordingly. Lately, there has been a surge of interest in mining biomedical literature. To this end, more and more information extraction (IE) systems using natural language processing (NLP) technologies have been developed for use in the biomedical field. Key biomedical IE tasks include named entity (NE) recognition (NER), such as the recognition of protein and gene names; and relation extraction, such as the extraction of protein-protein and gene-gene interactions.

NER identifies named entities from natural language texts and classifies them into specific classes according to a defined ontology or classification. In general, biomedical NEs do not follow any nomenclature and may comprise long compound words and short abbreviations. Some NEs contain various symbols and other spelling variations. On average, an NE has five synonyms (Tsai *et al*., 2006a), and it may belong to multiple categories intrinsically. Since biomedical language and vo-

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 243–246,
New York, June 2006.

cabulary are highly complex and evolving rapidly, Bio-NER is a very challenging problem, which raises a number of difficulties.

The other main focus of Bio-IE is relation extraction. Most systems only extract the relation targets (e.g., proteins, genes) and the verbs representing those relations, overlooking the many adverbial and prepositional phrases and words that describe location, manner, timing, condition, and extent. However, the information in such phrases may be important for precise definition and clarification of complex biological relations.

This problem can be tackled by using semantic role labeling (SRL) because it not only recognizes main roles, such as agents and objects, but also extracts adjunct roles such as location, manner, timing, condition, and extent. (Morarescu *et al.*, 2005) has demonstrated that full-parsing and SRL can improve the performance of relation extraction, resulting in an F-score increase of 15% (from 67% to 82%). This significant result leads us to surmise that SRL may also have potential for relation extraction in the biomedical domain. Unfortunately, no SRL system for the biomedical domain exists.

In this paper, we tackle the problems of both biomedical SRL and NER. Our contributions are (1) employing web lexicons and template-based post-processing to boost the performance of Bio-NER; (2) constructing a proposition bank on top of the popular biomedical GENIA treebank following the PropBank annotation scheme and developing a Biomedical SRL system. We adapt an SRL system trained the World Street Journal (WSJ) corpus to the biomedical domain. On adjunct arguments, especially those relevant to the biomedical domain, the performance is unsatisfactory. We, therefore, develop automatically generated templates for identifying these arguments.

2 Biomedical Named Entity Recognition

Our Bio-NER system uses the CRF model (Lafferty *et al.*, 2001), which has proven its effectiveness in several sequence tagging tasks.

2.1 Features and Post-Processing

Orthographical Features
In our experience, ALLCAPS, CAPSMIX, and INITCAP are more useful than others. The details are listed in (Tsai et al., 2006a).

Context Features
Words preceding or following the target word may be useful for determining its category. In our experience, a suitable window size is five.

Part-of-speech Features
Part-of-speech information is quite useful for identifying NEs. Verbs and prepositions usually indicate an NE's boundaries, whereas nouns not found in the dictionary are usually good candidates for named entities. Our experience indicates that five is also a suitable window size. The MBT POS tagger is used to provide POS information. We trained it on GENIA 3.02p and achieved 97.85% accuracy.

Word Shape Features
As NEs in the same category may look similar (e.g., IL-2 and IL-4), we have to find a simple way to normalize all similar words. According to our method, capitalized characters are all replaced by 'A', digits are all replaced by '0', non-English characters are replaced by '_' (underscore), and non-capitalized characters are replaced by 'a'. To further normalize these words, we reduce consecutive strings of identical characters to one character.

Affix Features
Some affixes can provide good clues for classifying named entities (e.g., "ase"). In our experience, an acceptable affix length is 3-5 characters.

Lexicon Features
Depending on the quality of a given dictionary, our system uses one of two different lexicon features to estimate the possibility of a token in a biomedical named entity. The first feature determines whether a token is part of a multi-word NE in the dictionary, while the second feature calculates the minimum distance between the given token and a dictionary. In our experience, the first feature is effective for a dictionary containing high-quality items, for example, human-curated protein dictionaries. The second feature is effective for a dictionary that has a large number of items that are not very accurate, for example, web or database lexicons. Details can be found in (Tsai et al., 2006a).

Post-Processing
We count the number of occurrences of a word x appearing in the rightmost position of all NEs in each category. Let the maximum occurrence be n,

and the corresponding category be *c*. The total number of occurrences of *x* in the rightmost position of an NE is *T*; c/T is the consistency rate of *x*. According to our analysis of the training set of the JNLPBA 2004 data, 75% of words have a consistency rate of over 95%. We record this 75% of words and their associated categories in a table. After testing, we crosscheck all the rightmost words of NEs found by our system against this table. If they match, we overwrite the NE categories with those from the table.

2.2 Experiments and Summary

We perform 10-fold cross validation on the GENIA V3.02 corpus (Kim *et al.*, 2003) to compare our CRF-based system with other biomedical NER systems. The experimental results are reported in Table 1. Our system outperforms other systems in protein names by an F-score of at least 2.6%. For DNA names, our performance is very close to that of the best system.

BioNER System	Protein	DNA
Our System (Tsai et al., 2006a)	78.4	66.3
HMM (Zhou et al., 2004)	75.8	63.3
Two Phase SVM (Lee et al., 2003)	70.6	66.4

Table 1. Performance of protein and DNA name recognition on the GENIA V3.02 corpus

We have made every effort to implement a variety of linguistic features in our system's CRF framework. Thanks to these features and the nature of CRF, our system outperforms state-of-the-art machine-learning-based systems, especially in the recognition of protein names.

Our system still has difficulty recognizing long, complicated NEs and coordinated NEs and distinguishing between overlapping NE classes, e.g., cell-line and cell-type. This is because biomedical texts have complicated sentence structures and involve more expert knowledge than texts from the general newswire domain. Since pure machine learning approaches cannot model long contextual phenomena well due to context window size limitations and data sparseness, we believe that template-based methods, which exploit long templates containing different levels of linguistic information, may be of help. Certain errors, such as incorrect boundary identification, are more tolerable if the main purpose is to discover relations between NEs (Tsai *et al.*, 2006c). We shall exploit more linguistic features, such as composite features and external features, in the future. However, machine leaning approaches suffer from a serious problem of annotation inconsistency, which confuses machine learning models and makes evaluation difficult. In order to reduce human annotation effort and alleviate the scarcity of available annotated corpora, we shall learn from web corpora to develop machine learning techniques in different biomedical domains.

3 Biomedical Semantic Role Labeling

In this section, we describe the main steps in building a biomedical SRL system: (1) create semantic roles for each biomedical verb; (2) construct a biomedical corpus, annotated with verbs and their corresponding semantic roles; (3) build an automatic semantic interpretation model, using the annotated text as a training corpus for machine learning. However, on adjunct arguments, especially on those highly relevant to the biomedical domain, such as AM-LOC (location), the performance is not satisfactory. We therefore develop a template generation method to create templates that are used as features for identifying these argument types.

3.1 Biomedical Proposition Bank -- BioProp

Our biomedical proposition bank, BioProp, is based on the GENIA Treebank (Yuka *et al.*, 2005), which is a 491-abstract corpus annotated with syntactic structures. The semantic annotation in BioProp is added to the proper constituents in a syntactic tree.

Basically, we adopt the definitions in PropBank (Palmer *et al.*, 2005). For the verbs not in PropBank, such as "phosphorylate", we define their framesets. Since the annotation is time-consuming, we adopt a semi-automatic approach. We adapt an SRL system trained on PropBank (Wall Street Journal corpus) to the biomedical domain. We first use this SRL system to automatically annotate our corpus, and then human annotators to double check the system's results. Therefore, human effort is greatly reduced.

3.2 Biomedical SRL System -- SEROW

Following (Punyakanok *et al*., 2004), we formulate SRL as a constituent-by-constituent (C-by-C) tagging problem. We use BioProp to train our biomedical SRL system, SEROW (Tsai *et al.*, 2006b), which uses a maximum entropy (ME) machine-learning model. We use the basic features described in (Xue & Palmer, 2004). In addition, we automatically generate templates which can be used to improve classification of biomedical argument types. The details of SEROW system are described in (Tsai *et al.*, 2005) and (Tsai *et al.*, 2006b).

3.3 Experiment and Summary

Our experimental results show that a newswire English SRL system that achieves an F-score of 86.29% can maintain an F-score of 64.64% when ported to the biomedical domain. By using SEROW, we can increase that F-score by 22.9%. Adding automatically generated template features further increases overall F-score by 0.47% and adjunct (AM) F-score by 1.57%, respectively.

4 Conclusion

NER and SRL are two key topics in biomedical NLP. For NER, we find broad linguistic features and integrate them into our CRF framework. Our system outperforms most machine learning-based systems, especially in the recognition of protein names (78.4% of F-score). In the future, templates that can match long contextual relations and coordinated NEs may be applied to NER post-processing. Web corpora may also be used to enhance unknown NE detection. In Bio-SRL, our contribution is threefold. First, we construct a biomedical proposition bank, BioProp, on top of the popular biomedical GENIA treebank following the PropBank annotation scheme. We employ semi-automatic annotation using an SRL system trained on PropBank thereby significantly reducing annotation effort. Second, we construct SEROW, which uses BioProp as its training corpus. Thirdly, we develop a method to automatically generate templates that can boost overall performance, especially on location, manner, adverb, and temporal arguments. In the future, we will expand BioProp to include more biomedical verbs and will also integrate a parser into SEROW.

References

Kim, J.-D., Ohta, T., Teteisi, Y., & Tsujii, J. i. (2003). Genia corpus - a semantically annotated corpus for bio-textmining. Bioinformatics, 19(suppl. 1).

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. Paper presented at the ICML-01.

Lee, K.-J., Hwang, Y.-S., & Rim, H.-C. (2003). Two phase biomedical ne recognition based on svms. Paper presented at the ACL-03 Workshop on Natural Language Processing in Biomedicine.

Morarescu, P., Bejan, C., & Harabagiu, S. (2005). Shallow semantics for relation extraction. Paper presented at the IJCAI-05.

Palmer, M., Gildea, D., & Kingsbury, P. (2005). The proposition bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1).

Punyakanok, V., Roth, D., Yih, W., & Zimak, D. (2004). Semantic role labeling via integer linear programming inference. Paper presented at the 20th International Conference on Computational Linguistics (COLING-04).

Tsai, R. T.-H., Chou, W.-C., Wu, S.-H., Sung, T.-Y., Hsiang, J., & Hsu, W.-L. (2006a). Integrating linguistic knowledge into a conditional random field framework to identify biomedical named entities. Expert Systems with Applications, 30(1), 117-128.

Tsai, R. T.-H., Lin, W.-C. C. Y.-C., Ku, W., Su, Y.-S., Sung, T.-Y., & Hsu, W.-L. (2006b). Serow: Adapting semantic role labeling for biomedical verbs: An exponential model coupled with adapting semantic role labeling for biomedical verbs: An exponential model coupled with automatically generated template features. To appear in BioNLP-2006.

Tsai, R. T.-H., Wu, C.-W., Lin, Y.-C., & Hsu, W.-L. (2005). Exploiting full parsing information to label semantic roles using an ensemble of me and svm via integer linear programming. Paper presented at the CoNLL-2005.

Tsai, R. T.-H., Wu, S.-H., Chou, W.-C., Lin, Y.-C., He, D., Hsiang, J., et al. (2006c). Various criteria in the evaluation of biomedical named entity recognition. BMC Bioinformatics, 7(92).

Xue, N., & Palmer, M. (2004). Calibrating features for semantic role labeling. Paper presented at the EMNLP 2004.

Yuka, T., Yakushiji, A., Ohta, T., & Tsujii, J. (2005). Syntax annotation for the genia corpus.

Zhou, G., Zhang, J., Su, J., Shen, D., & Tan, C. (2004). Recognizing names in biomedical texts: A machine learning approach. Bioinformatics, 20, 1178-1190.

Supporting Multiple Information-Seeking Strategies in a Single System Framework

Xiaojun Yuan
School of Communication, Information and Library Studies
Rutgers, The State University of New Jersey
New Brunswick, NJ 08901, USA
yuanxj@rci.rutgers.edu

Abstract

This study investigates the support of multiple information seeking strategies (ISSs) within a single system, and the relation between varieties of ISSs and system design. It proposes to construct and evaluate an interactive information retrieval system which can adaptively support multiple ISSs, and allow change from one ISS to another within an ISS space. It is conducted in a series of steps: iterative designing -evaluating of several systems supporting different ISSs; specifying an interaction structure for multiple ISSs; and, implementing and evaluating a dynamically adaptive system supporting multiple ISSs. The study aims to make a contribution to interactive information retrieval drawing attention to user interface design, and to HCI, in integration of multiple support techniques within a single system framework.

Keywords

Information-seeking strategy, interaction structure, user interface design, evaluation, information retrieval

1 Introduction

Traditional information retrieval (IR) systems ask people to specify queries using terms to select documents from the selected databases. Current research indicates some problems of such systems. (1) Human information-seeking behavior is more complicated than just query formulation and term selection. For instance, people need to get an idea of which domain or genre of information they need, then familiarize themselves with the content and structure of various databases. They need to learn about a domain of interest and extend their knowledge of this domain so that they can formulate more effective queries. They need to change their searching and seeking strategies to improve their queries. (2) Human information-seeking behaviors are not discrete processes. These behaviors interact with one another during information-seeking processes (Belkin, 1996). (3) Interaction, not representation or comparison, is the central process of IR (Belkin, 1993). (4) Users with different goals need to use different information-seeking strategies (ISSs) to conduct searches. However, traditional IR systems only support one ISS, that is, formulating queries using terms to select documents from some databases (Belkin, 1993, 1995; Xie, 2000). (5) It is known that different ISSs can be optimally supported by different combinations of IR techniques (Oddy, 1977). The existing diversity of ISSs indicates that a system which provides good support for one ISS is unlikely to provide good support for the others. A system trying to support all ISSs with only one technique will be able to support them at only mediocre levels.

Therefore, the goal of the research is to design an interactive IR system which incorporates different IR techniques to adaptively support different ISSs. Specifically, our solution to these problems focuses mainly on the following two goals.

(1) Constructing a single IR system in which support techniques are explicitly represented so that it is possible to shift from one combination of support techniques to another in real time, and appropriate support techniques are suggested to the user by the system.
(2) Evaluating the effectiveness and usability of the system within controlled experiments.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 247–249,
New York, June 2006.

2 Research Problems

We aim to investigate the following research problems:
(1) Implementing and evaluating several systems which are tailored to scanning or searching.
(2) Developing a structure for guiding and controlling sequences of different support techniques.
(3) Constructing and evaluating a single system which supports scanning and searching through integration of different support techniques within a single system framework.

3 Methodology

3.1 Research Problem 1

SYSTEMS

Using the Lemur toolkit (LEMUR), we implemented and evaluated several different prototype IR systems designed to support scanning (situation 1) or searching (situation 2). Table 1 describes the tasks, as well as features and support techniques for each system.

HYPOTHESES

Hypothesis 1: The system summarizing each database is more effective in supporting scanning tasks than the baseline system which provides a ranked list of documents with descriptions about which databases these documents are in. (E1.1/B1.1, situation1-task1)

Hypothesis 2: The system providing table of contents navigation is more effective in supporting scanning tasks than the baseline system which lists ranked paragraphs. (E1.2/B1.2, situation1-task2)

Hypothesis 3: The system presenting clustered retrieval results is more effective in supporting searching tasks than the baseline system which presents a ranked list of retrieval results. (E2.1/B2.1, situation2-task1)

Hypothesis 4: The system supporting fielded queries is more effective in supporting searching tasks than the baseline system which provides a generic query search. (E2.2/B2.2, situation2-task2)

Situations	Tasks	Experimental Systems	Baseline Systems	Support Techniques
1 Scanning	1 Identify best databases	E1.1 Alphabetically ordered databases (showing summary for each)	B1.1 Ranked documents (showing name of the database with the document)	Summary of each database
	2 Find comments or quotations from an electronic book	E1.2 Table of Contents navigation within documents	B1.2 Ranked paragraphs	Table of contents navigation
2 Searching	1 Find relevant documents	E2.1 Ranked clusters	B2.1 Ranked documents	Clustered retrieval results
	2 Find the name of an electronic book	E2.2 Field search	B2.2 Generic query search	Fielded query

Table 1. Situations, Tasks and Systems

EXPERIMENTAL DESIGN

Participants conducted four searches on four different topics that are suitable for scanning or searching. This is a within-subjects design. Each subject searched half of the topics in one system, then half of the topics in the other system. Within the topic block, the topic order was randomly assigned. No two subjects used the same order of topics and the same order of systems. The experiments were replicated by exchanging the order of the systems.

TEXT COLLECTIONS

There are two text collections: one is TREC HARD 2004 collection (HARD) which is suitable for situation1-task1 and situation2-task1, the other is a book database which is good for situation1-task2 and situation2-task2. This database is composed of books downloaded from Project Gutenberg (Gutenberg).

TASKS

In this study, we used the simulated work task situation model (Borlund, 1997) to make subjects' behavior as true-to-life as possible, hoping this will

make our results robust. Here is an example for situation1-task1.

> Topic: As a graduate student, you are asked to write an essay about air pollution for one of your courses. You are supposed to get information you need from a system that is composed of several databases. Each database has lots of documents on a variety of topics. You believe it would be interesting to discover factors that cause air pollution, but you have no idea which databases are good on this topic.
> Task: Please find out which databases are good for this particular topic, and rank the databases in order of likelihood of being good. Put your answer in the given space.

3.2 Research Problem 2 (Future Work)

In order to guide the presentation of specific support techniques during the information seeking process, we need to specify an interaction structure. This interaction structure is equivalent to a dialogue manager, and can be used to control the interactions between the system and the user. We will employ the idea of interaction structure developed in the MERIT system (Belkin, 1995). This structure models human-computer interaction as dialogues and particular dialogue structures are associated with different ISSs. This structure will be incorporated into the system at the user interface level and act as the dialogue manager.

3.3 Research Problem 3 (Future Work)

SYSTEM

The integrated system will allow the user to use a variety of ISSs and to seamlessly switch from one ISS to another in the information-seeking process. The user will be able to choose which ISS to use at any time. ISSs will be classified according to the goal of the interaction, the topic or task, and the information-seeking stage, etc. The system should be able to suggest to the user appropriate ISSs at the appropriate times, given the current state of the information- seeking process.

HYPOTHESIS

Hypothesis 5: The integrated system purposely designed for supporting both scanning and searching is more effective in supporting tasks requiring scanning and searching than the generic baseline system.

EXPERIMENTAL DESIGN

This will be a within-subject experimental design. The subjects will search the integrated system and then the baseline system. The experiment will be replicated by changing the order of the systems.

4 Conclusion

Our aim is to contribute to the field of interactive information retrieval drawing attention to the user interface design and HCI. The systems in research problem 1 have been implemented and the user studies were conducted. Future work will focus on the interaction structure and construction and testing of the integrated system. Through this we hope to improve information retrieval and human-computer interaction.

References

Belkin, N. J., Marchetti, P. G., & Cool, C. BRAQUE: Design of an interface to support user interaction in information retrieval. 1993. *Information Processing & management,* 29(3): 325-344.

Belkin, N.J., Cool, C., Stein, A., Theil, U., Cases, scripts and information seeking strategies: on the design of interactive information retrieval systems. 1995. *Expert Systems with Applications*, 9(3): 379-395.

Belkin, N. J. Intelligent Information Retrieval: Whose Intelligence? 1996. In *Proceedings of the Fifth International Symposium for Information Science (ISI-96)*, 25-31.

Borlund,P. & Ingwerson, P. The development of a method for the evaluation of interactive information retrieval systems. 1997. *Journal of Documentation,* 53(3): 225-250.

GUTENBERG. http://www.gutenberg.org/

HARD. http://projects.ldc.upenn.edu/HARD/

LEMUR. http://www.lemurproject.org/

Oddy, R.N. Information retrieval through man-machine dialogue. 1977. *Journal of Documentation,*33(1): 1-14.

Xie, H. Shifts of interactive intentions and information-seeking strategies in interactive information retrieval. 2000. *Journal of the American Society for Information Science*, 51(9): 841-857.

Author Index