The impact of language models and loss functions on repair disfluency detection

Simon Zwarts and Mark Johnson
Centre for Language Technology
Macquarie University
{simon.zwarts|mark.johnson|}@mq.edu.au

Abstract

Unrehearsed spoken language often contains disfluencies. In order to correctly interpret a spoken utterance, any such disfluencies must be identified and removed or otherwise dealt with. Operating on transcripts of speech which contain disfluencies, we study the effect of language model and loss function on the performance of a linear reranker that rescores the 25-best output of a noisy-channel model. We show that language models trained on large amounts of non-speech data improve performance more than a language model trained on a more modest amount of speech data, and that optimising f-score rather than log loss improves disfluency detection performance.

Our approach uses a log-linear reranker, operating on the top n analyses of a noisy channel model. We use large language models, introduce new features into this reranker and examine different optimisation strategies. We obtain a disfluency detection f-scores of 0.838 which improves upon the current state-of-the-art.

1 Introduction

Most spontaneous speech contains disfluencies such as partial words, filled pauses (e.g., "uh", "um", "huh"), explicit editing terms (e.g., "I mean"), parenthetical asides and repairs. Of these, repairs pose particularly difficult problems for parsing and related Natural Language Processing (NLP) tasks. This paper presents a model of disfluency detection based on the noisy channel framework, which specifically targets the repair disfluencies. By combining language models and using an appropriate loss function in a log-linear reranker we are able to achieve f-scores which are higher than previously reported.

Often in natural language processing algorithms, more data is more important than better algorithms (Brill and Banko, 2001). It is this insight that drives the first part of the work described in this paper. This paper investigates how we can use language models trained on large corpora to increase repair detection accuracy performance.

There are three main innovations in this paper. First, we investigate the use of a variety of language models trained from text or speech corpora of various genres and sizes. The largest available language models are based on written text: we investigate the effect of written text language models as opposed to language models based on speech transcripts. Second, we develop a new set of reranker features explicitly designed to capture important properties of speech repairs. Many of these features are lexically grounded and provide a large performance increase. Third, we utilise a loss function, approximate expected f-score, that explicitly targets the asymmetric evaluation metrics used in the disfluency detection task. We explain how to optimise this loss function, and show that this leads to a marked improvement in disfluency detection. This is consistent with Jansche (2005) and Smith and Eisner (2006), who observed similar improvements when using approximate f-score loss for other problems. Similarly we introduce a loss function based on the edit-f-score in our domain.

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pages 703–711,
Portland, Oregon, June 19-24, 2011.

Together, these three improvements are enough to boost detection performance to a higher f-score than previously reported in literature. Zhang et al. (2006) investigate the use of 'ultra large feature spaces' as an aid for disfluency detection. Using over 19 million features, they report a final f-score in this task of 0.820. Operating on the same body of text (Switchboard), our work leads to an f-score of 0.838, this is a 9% relative improvement in residual f-score.

The remainder of this paper is structured as follows. First in Section 2 we describe related work. Then in Section 3 we present some background on disfluencies and their structure. Section 4 describes appropriate evaluation techniques. In Section 5 we describe the noisy channel model we are using. The next three sections describe the new additions: Section 6 describe the corpora used for language models, Section 7 describes features used in the log-linear model employed by the reranker and Section 8 describes appropriate loss functions which are critical for our approach. We evaluate the new model in Section 9. Section 10 draws up a conclusion.

2 Related work

A number of different techniques have been proposed for automatic disfluency detection. Schuler et al. (2010) propose a Hierarchical Hidden Markov Model approach; this is a statistical approach which builds up a syntactic analysis of the sentence and marks those subtrees which it considers to be made up of disfluent material. Although they are interested not only in disfluency but also a syntactic analysis of the utterance, including the disfluencies being analysed, their model's final f-score for disfluency detection is lower than that of other models.

Snover et al. (2004) investigate the use of purely lexical features combined with part-of-speech tags to detect disfluencies. This approach is compared to approaches which use primarily prosodic cues, and appears to perform equally well. However, the authors note that this model finds it difficult to identify disfluencies which by themselves are very fluent. As we will see later, the individual components of a disfluency do not have to be disfluent by themselves. This can occur when a speaker edits her speech for meaning-related reasons, rather than errors that arise from performance. The edit repairs which are the focus of our work typically have this characteristic.

Noisy channel models have done well on the disfluency detection task in the past; the work of Johnson and Charniak (2004) first explores such an approach. Johnson et al. (2004) adds some hand-written rules to the noisy channel model and use a maximum entropy approach, providing results comparable to Zhang et al. (2006), which are state-of-the art results.

Kahn et al. (2005) investigated the role of prosodic cues in disfluency detection, although the main focus of their work was accurately recovering and parsing a fluent version of the sentence. They report a 0.782 f-score for disfluency detection.

3 Speech Disfluencies

We follow the definitions of Shriberg (1994) regarding speech disfluencies. She identifies and defines three distinct parts of a speech disfluency, referred to as the *reparandum*, the *interregnum* and the *repair*. Consider the following utterance:

$$\begin{array}{c} \textit{I want a flight } \overbrace{\textit{to Boston,}}^{\text{reparandum}} \\ \underbrace{\textit{uh, I mean}}_{\text{interregnum}} \underbrace{\textit{to Denver}}_{\text{repair}} \; on\ Friday \end{array} \tag{1}$$

The reparandum *to Boston* is the part of the utterance that is 'edited out'; the interregnum *uh, I mean* is a filled pause, which need not always be present; and the repair *to Denver* replaces the reparandum.

Shriberg and Stolcke (1998) studied the location and distribution of repairs in the Switchboard corpus (Godfrey and Holliman, 1997), the primary corpus for speech disfluency research, but did not propose an actual model of repairs. They found that the overall distribution of speech disfluencies in a large corpus can be fit well by a model that uses only information on a very local level. Our model, as explained in section 5, follows from this observation.

As our domain of interest we use the Switchboard corpus. This is a large corpus consisting of transcribed telephone conversations between two partners. In the Treebank III (Marcus et al., 1999) corpus there is annotation available for the Switchboard corpus, which annotates which parts of utterances are in a reparandum, interregnum or repair.

4 Evaluation metrics for disfluency detection systems

Disfluency detection systems like the one described here identify a subset of the word tokens in each transcribed utterance as "edited" or disfluent. Perhaps the simplest way to evaluate such systems is to calculate the accuracy of labelling they produce, i.e., the fraction of words that are correctly labelled (i.e., either "edited" or "not edited"). However, as Charniak and Johnson (2001) observe, because only 5.9% of words in the Switchboard corpus are "edited", the trivial baseline classifier which assigns all words the "not edited" label achieves a labelling accuracy of 94.1%.

Because the labelling accuracy of the trivial baseline classifier is so high, it is standard to use a different evaluation metric that focuses more on the detection of "edited" words. We follow Charniak and Johnson (2001) and report the f-score of our disfluency detection system. The f-score f is:

$$f \quad = \quad \frac{2c}{g+e} \qquad (2)$$

where g is the number of "edited" words in the gold test corpus, e is the number of "edited" words proposed by the system on that corpus, and c is the number of the "edited" words proposed by the system that are in fact correct. A perfect classifier which correctly labels every word achieves an f-score of 1, while the trivial baseline classifiers which label every word as "edited" or "not edited" respectively achieve a very low f-score.

Informally, the f-score metric focuses more on the "edited" words than it does on the "not edited" words. As we will see in section 8, this has implications for the choice of loss function used to train the classifier.

5 Noisy Channel Model

Following Johnson and Charniak (2004), we use a noisy channel model to propose a 25-best list of possible speech disfluency analyses. The choice of this model is driven by the observation that the repairs frequently seem to be a "rough copy" of the reparandum, often incorporating the same or very similar words in roughly the same word order. That is, they seem to involve "crossed" dependencies between the reparandum and the repair. Example (3) shows the crossing dependencies. As this example also shows, the repair often contains many of the same words that appear in the reparandum. In fact, in our Switchboard training corpus we found that 62reparandum also appeared in the associated repair,

$$\underbrace{\textit{to Boston}}_{\text{reparandum}}\ \underbrace{\textit{uh, I mean,}}_{\text{interregnum}}\ \underbrace{\textit{to Denver}}_{\text{repair}} \qquad (3)$$

5.1 Informal Description

Given an observed sentence Y we wish to find the most likely source sentence $\hat{X}$, where

$$\hat{X} = \operatorname*{argmax}_{X} P(Y|X)P(X) \qquad (4)$$

In our model the unobserved X is a substring of the complete utterance Y.

Noisy-channel models are used in a similar way in statistical speech recognition and machine translation. The language model assigns a probability $P(X)$ to the string X, which is a substring of the observed utterance Y. The channel model $P(Y|X)$ generates the utterance Y, which is a potentially disfluent version of the source sentence X. A repair can potentially begin before any word of X. When a repair has begun, the channel model incrementally processes the succeeding words from the start of the repair. Before each succeeding word either the repair can end or else a sequence of words can be inserted in the reparandum. At the end of each repair, a (possibly null) interregnum is appended to the reparandum.

We will look at these two components in the next two Sections in more detail.

5.2 Language Model

Informally, the task of language model component of the noisy channel model is to assess fluency of the sentence with disfluency removed. Ideally we would like to have a model which assigns a very high probability to disfluency-free utterances and a lower probability to utterances still containing disfluencies. For computational complexity reasons, as described in the next section, inside the noisy channel model we use a bigram language model. This bigram language model is trained on the fluent version of the Switchboard corpus (training section).

We realise that a bigram model might not be able to capture more complex language behaviour. This motivates our investigation of a range of additional language models, which are used to define features used in the log-linear reranker as described below.

5.3 Channel Model

The intuition motivating the channel model design is that the words inserted into the reparandum are very closely related to those in the repair. Indeed, in our training data we find that 62% of the words in the reparandum are exact copies of words in the repair; this identity is strong evidence of a repair. The channel model is designed so that exact copy reparandum words will have high probability.

Because these repair structures can involve an unbounded number of crossed dependencies, they cannot be described by a context-free or finite-state grammar. This motivates the use of a more expressive formalism to describe these repair structures.

We assume that X is a substring of Y, i.e., that the source sentence can be obtained by deleting words from Y, so for a fixed observed utterance Y there are only a finite number of possible source sentences. However, the number of possible source sentences, X, grows exponentially with the length of Y, so exhaustive search is infeasible. Tree Adjoining Grammars (TAG) provide a systematic way of formalising the channel model, and their polynomial-time dynamic programming parsing algorithms can be used to search for likely repairs, at least when used with simple language models like a bigram language model. In this paper we first identify the 25 most likely analyses of each sentence using the TAG channel model together with a bigram language model.

Further details of the noisy channel model can be found in Johnson and Charniak (2004).

5.4 Reranker

To improve performance over the standard noisy channel model we use a reranker, as previously suggest by Johnson and Charniak (2004). We rerank a 25-best list of analyses. This choice is motivated by an oracle experiment we performed, probing for the location of the best analysis in a 100-best list. This experiment shows that in 99.5% of the cases the best analysis is located within the first 25, and indicates that an f-score of 0.958 should be achievable as the upper bound on a model using the first 25 best analyses. We therefore use the top 25 analyses from the noisy channel model in the remainder of this paper and use a reranker to choose the most suitable candidate among these.

6 Corpora for language modelling

We would like to use additional data to model the fluent part of spoken language. However, the Switchboard corpus is one of the largest widely-available disfluency-annotated speech corpora. It is reasonable to believe that for effective disfluency detection Switchboard is not large enough and more text can provide better analyses. Schwartz et al. (1994), although not focusing on disfluency detection, show that using written language data for modelling spoken language can improve performance. We turn to three other bodies of text and investigate the use of these corpora for our task, disfluency detection. We will describe these corpora in detail here.

The predictions made by several language models are likely to be strongly correlated, even if the language models are trained on different corpora. This motivates the choice for log-linear learners, which are built to handle features which are not necessarily independent. We incorporate information from the external language models by defining a reranker feature for each external language model. The value of this feature is the log probability assigned by the language model to the candidate underlying fluent substring X

For each of our corpora (including Switchboard) we built a 4-gram language model with Kneser-Ney smoothing (Kneser and Ney, 1995). For each analysis we calculate the probability under that language model for the candidate underlying fluent substring X. We use this log probability as a feature in the reranker. We use the SRILM toolkit (Stolcke, 2002) both for estimating the model from the training corpus as well as for computing the probabilities of the underlying fluent sentences X of the different analysis.

As previously described, **Switchboard** is our primary corpus for our model. The language model part of the noisy channel model already uses a bigram language model based on Switchboard, but in the reranker we would like to also use 4-grams for reranking. Directly using Switchboard to build a 4-gram language model is slightly problematic. When we use the training data of Switchboard both for language fluency prediction and the same training data also for the loss function, the reranker will overestimate the weight associated with the feature derived from the Switchboard language model, since the fluent sentence itself is part of the language model training data. We solve this by dividing the Switchboard training data into 20 folds. For each fold we use the 19 other folds to construct a language model and then score the utterance in this fold with that language model.

The largest widely-available corpus for language modelling is the **Web 1T 5-gram** corpus (Brants and Franz, 2006). This data set, collected by Google Inc., contains English word n-grams and their observed frequency counts. Frequency counts are produced from this billion-token corpus of web text. Because of the noise[1] present in this corpus there is an ongoing debate in the scientific community of the use of this corpus for serious language modelling.

The **Gigaword** Corpus (Graff and Cieri, 2003) is a large body of newswire text. The corpus contains $1.6 \cdot 10^9$ tokens, however fluent newswire text is not necessarily of the same domain as disfluency removed speech.

The **Fisher** corpora Part I (David et al., 2004) and Part II (David et al., 2005) are large bodies of transcribed text. Unlike Switchboard there is no disfluency annotation available for Fisher. Together the two Fisher corpora consist of $2.2 \cdot 10^7$ tokens.

7 Features

The log-linear reranker, which rescores the 25-best lists produced by the noisy-channel model, can also include additional features besides the noisy-channel log probabilities. As we show below, these additional features can make a substantial improvement to disfluency detection performance. Our reranker incorporates two kinds of features. The first are log-probabilities of various scores computed by the noisy-channel model and the external language models. We only include features which occur at least 5 times in our training data.

The noisy channel and language model features consist of:

1. LMP: 4 features indicating the probabilities of the underlying fluent sentences under the language models, as discussed in the previous section.

2. NCLogP: The Log Probability of the entire noisy channel model. Since by itself the noisy channel model is already doing a very good job, we do not want this information to be lost.

3. LogFom: This feature is the log of the “figure of merit” used to guide search in the noisy channel model when it is producing the 25-best list for the reranker. The log figure of merit is the sum of the log language model probability and the log channel model probability plus 1.5 times the number of edits in the sentence. This feature is redundant, i.e., it is a linear combination of other features available to the reranker model: we include it here so the reranker has direct access to all of the features used by the noisy channel model.

4. NCTransOdd: We include as a feature parts of the noisy channel model itself, i.e. the channel model probability. We do this so that the task to choosing appropriate weights of the channel model and language model can be moved from the noisy channel model to the log-linear optimisation algorithm.

The boolean indicator features consist of the following 3 groups of features operating on words and their edit status; the latter indicated by one of three possible flags: _ when the word is not part of a disfluency or E when it is part of the reparandum or I when it is part of the interregnum.

1. CopyFlags_X_Y: When there is an exact copy in the input text of length X ($1 \leq X \leq 3$) and the gap between the copies is Y ($0 \leq Y \leq 3$) this feature is the sequence of flags covering the two copies. Example: `CopyFlags_1_0 (E`

[1]We do not mean speech disfluencies here, but noise in web-text; web-text is often poorly written and unedited text.

_) records a feature when two identical words are present, directly consecutive and the first one is part of a disfluency (**E**dited) while the second one is not. There are 745 different instances of these features.

2. WordsFlags_L_n_R: This feature records the immediate area around an n-gram ($n \leq 3$). L denotes how many flags to the left and R ($0 \leq R \leq 1$) how many to the right are includes in this feature (Both L and R range over 0 and 1). Example: `WordsFlags_1_1_0 (need _)` is a feature that fires when a fluent word is followed by the word 'need' (one flag to the left, none to the right). There are 256808 of these features present.

3. SentenceEdgeFlags_B_L: This feature indicates the location of a disfluency in an utterance. The Boolean B indicates whether this features records sentence initial or sentence final behaviour, L ($1 \leq L \leq 3$) records the length of the flags. Example `SentenceEdgeFlags_1_1 (I)` is a feature recording whether a sentence ends on an interregnum. There are 22 of these features present.

We give the following analysis as an example:

but E *but* _ *that* _ *does* _ *n't* _ *work* _

The language model features are the probability calculated over the fluent part. NCLogP, LogFom and NCTransOdd are present with their associated value. The following binary flags are present:

```
CopyFlags_1_0 (E _)
WordsFlags:0:1:0 (but E)
WordsFlags:0:1:0 (but _)
WordsFlags:1:1:0 (E but _)
WordsFlags:1:1:0 (_ that _)
WordsFlags:0:2:0 (but E but _) etc.[2]
SentenceEdgeFlags:0:1 (E)
SentenceEdgeFlags:0:2 (E _)
SentenceEdgeFlags:0:3 (E _ _)
```

These three kinds of boolean indicator features together constitute the *extended feature set*.

[2]An exhaustive list here would be too verbose.

8 Loss functions for reranker training

We formalise the reranker training procedure as follows. We are given a training corpus T containing information about n possibly disfluent sentences. For the ith sentence T specifies the sequence of words x_i, a set $\mathcal{Y}_i$ of 25-best candidate "edited" labellings produced by the noisy channel model, as well as the correct "edited" labelling $y_i^\star \in \mathcal{Y}_i$.[3]

We are also given a vector $\mathbf{f} = (f_1, \ldots, f_m)$ of *feature functions*, where each f_j maps a word sequence x and an "edit" labelling y for x to a real value $f_j(x, y)$. Abusing notation somewhat, we write $\mathbf{f}(x, y) = (f_1(x, y), \ldots, f_m(x, y))$. We interpret a vector $\mathbf{w} = (w_1, \ldots, w_m)$ of *feature weights* as defining a conditional probability distribution over a candidate set $\mathcal{Y}$ of "edited" labellings for a string x as follows:

$$\mathrm{P}_{\mathbf{w}}(y \mid x, \mathcal{Y}) = \frac{\exp(\mathbf{w} \cdot \mathbf{f}(x, y))}{\sum_{y' \in \mathcal{Y}} \exp(\mathbf{w} \cdot \mathbf{f}(x, y'))}$$

We estimate the feature weights $\mathbf{w}$ from the training data T by finding a feature weight vector $\widehat{\mathbf{w}}$ that optimises a regularised objective function:

$$\widehat{\mathbf{w}} = \operatorname*{argmin}_{\mathbf{w}} L_T(\mathbf{w}) + \alpha \sum_{j=1}^{m} w_j^2$$

Here α is the *regulariser weight* and L_T is a *loss function*. We investigate two different loss functions in this paper. $\mathit{LogLoss}$ is the negative log conditional likelihood of the training data:

$$\mathit{LogLoss}_T(\mathbf{w}) = \sum_{i=1}^{m} -\log \mathrm{P}(y_i^\star \mid x_i, \mathcal{Y}_i)$$

Optimising $\mathit{LogLoss}$ finds the $\widehat{\mathbf{w}}$ that define (regularised) conditional Maximum Entropy models.

It turns out that optimising $\mathit{LogLoss}$ yields suboptimal weight vectors $\widehat{\mathbf{w}}$ here. $\mathit{LogLoss}$ is a symmetric loss function (i.e., each mistake is equally weighted), while our f-score evaluation metric weights "edited" labels more highly, as explained in section 4. Because our data is so skewed (i.e., "edited" words are comparatively infrequent), we

[3]In the situation where the true "edited" labelling does not appear in the 25-best list $\mathcal{Y}_i$ produced by the noisy-channel model, we choose $y_i^\star$ to be a labelling in $\mathcal{Y}_i$ closest to the true labelling.

can improve performance by using an asymmetric loss function.

Inspired by our evaluation metric, we devised an *approximate expected f-score loss function* $FLoss$.

$$FLoss_T(\mathbf{w}) \;=\; 1 - \frac{2\mathrm{E}_{\mathbf{w}}[c]}{g + \mathrm{E}_{\mathbf{w}}[e]}$$

This approximation assumes that the expectations approximately distribute over the division: see Jansche (2005) and Smith and Eisner (2006) for other approximations to expected f-score and methods for optimising them. We experimented with other asymmetric loss functions (e.g., the expected error rate) and found that they gave very similar results.

An advantage of $FLoss$ is that it and its derivatives with respect to $\mathbf{w}$ (which are required for numerical optimisation) are easy to calculate exactly. For example, the expected number of correct "edited" words is:

$$\begin{aligned}\mathrm{E}_{\mathbf{w}}[c] &= \sum_{i=1}^{n} \mathrm{E}_{\mathbf{w}}[c_{y_i^\star} \mid \mathcal{Y}_i], \text{ where:}\\ \mathrm{E}_{\mathbf{w}}[c_{y_i^\star} \mid \mathcal{Y}_i] &= \sum_{y \in \mathcal{Y}_i} c_{y_i^\star}(y)\, \mathrm{P}_{\mathbf{w}}(y \mid x_i, \mathcal{Y}_i)\end{aligned}$$

and $c_{y^\star}(y)$ is the number of correct "edited" labels in y given the gold labelling $y^\star$. The derivatives of $FLoss$ are:

$$\frac{\partial FLoss_T}{\partial w_j}(\mathbf{w}) \;=\; \frac{1}{g + \mathrm{E}_{\mathbf{w}}[e]} \left(FLoss_T(\mathbf{w}) \frac{\partial \mathrm{E}_{\mathbf{w}}[e]}{\partial w_j} - 2 \frac{\partial \mathrm{E}_{\mathbf{w}}[c]}{\partial w_j} \right)$$

where:

$$\frac{\partial \mathrm{E}_{\mathbf{w}}[c]}{\partial w_j} \;=\; \sum_{i=1}^{n} \frac{\partial \mathrm{E}_{\mathbf{w}}[c_{y_i^\star} \mid x_i, \mathcal{Y}_i]}{\partial w_j}$$

$$\frac{\partial \mathrm{E}_{\mathbf{w}}[c_{y^\star} \mid x, \mathcal{Y}]}{\partial w_j} \;=\; \mathrm{E}_{\mathbf{w}}[f_j c_{y^\star} \mid x, \mathcal{Y}] - \mathrm{E}_{\mathbf{w}}[f_j \mid x, \mathcal{Y}]\, \mathrm{E}_{\mathbf{w}}[c_{y^\star} \mid x, \mathcal{Y}].$$

$\partial \mathrm{E}[e] / \partial w_j$ is given by a similar formula.

9 Results

We follow Charniak and Johnson (2001) and split the corpus into main training data, held-out training data and test data as follows: main training consisted of all sw[23]$*$.dps files, held-out training consisted of all sw4[5-9]$*$.dps files and test consisted of all sw4[0-1]$*$.dps files. However, we follow (Johnson and Charniak, 2004) in deleting all partial words and punctuation from the training and test data (they argued that this is more realistic in a speech processing application).

Table 1 shows the results for the different models on held-out data. To avoid over-fitting on the test data, we present the f-scores over held-out training data instead of test data. We used the held-out data to select the best-performing set of reranker features, which consisted of features for all of the language models plus the extended (i.e., indicator) features, and used this model to analyse the test data. The f-score of this model on test data was 0.838. In this table, the set of *Extended Features* is defined as all the boolean features as described in Section 7.

We first observe that adding different external language models does increase the final score. The difference between the external language models is relatively small, although the differences in choice are several orders of magnitude. Despite the putative noise in the corpus, a language model built on Google's Web1T data seems to perform very well. Only the model where Switchboard 4-grams are used scores slightly lower, we explain this because the internal bigram model of the noisy channel model is already trained on Switchboard and so this model adds less new information to the reranker than the other models do.

Including additional features to describe the problem space is very productive. Indeed the best performing model is the model which has all extended features and all language model features. The differences among the different language models when extended features are present are relatively small. We assume that much of the information expressed in the language models overlaps with the lexical features.

We find that using a loss function related to our evaluation metric, rather than optimising $LogLoss$, consistently improves edit-word f-score. The standard $LogLoss$ function, which estimates the "maximum entropy" model, consistently performs worse than the loss function minimising expected errors.

The best performing model (Base + Ext. Feat. + All LM, using expected f-score loss) scores an f-score of **0.838** on **test data**. The results as indicated by the f-score outperform state-of-the-art models re-

Model	**F-score**	
Base (noisy channel, no reranking)	0.756	
Model	**log loss**	**expected f-score loss**
Base + Switchboard	0.776	0.791
Base + Fisher	0.771	0.797
Base + Gigaword	0.777	0.797
Base + Web1T	0.781	0.798
Base + Ext. Feat.	0.824	0.827
Base + Ext. Feat. + Switchboard	0.827	0.828
Base + Ext. Feat. + Fisher	0.841	0.856
Base + Ext. Feat. + Gigaword	0.843	0.852
Base + Ext. Feat. + Web1T	0.843	0.850
Base + Ext. Feat. + All LM	0.841	**0.857**

Table 1: Edited word detection f-score on held-out data for a variety of language models and loss functions

ported in literature operating on identical data, even though we use vastly less features than other do.

10 Conclusion and Future work

We have described a disfluency detection algorithm which we believe improves upon current state-of-the-art competitors. This model is based on a noisy channel model which scores putative analyses with a language model; its channel model is inspired by the observation that reparandum and repair are often very similar. As Johnson and Charniak (2004) noted, although this model performs well, a log-linear reranker can be used to increase performance.

We built language models from a variety of speech and non-speech corpora, and examine the effect they have on disfluency detection. We use language models derived from different larger corpora effectively in a maximum reranker setting. We show that the actual choice for a language model seems to be less relevant and newswire text can be used equally well for modelling fluent speech.

We describe different features to improve disfluency detection even further. Especially these features seem to boost performance significantly.

Finally we investigate the effect of different loss functions. We observe that using a loss function directly optimising our interest yields a performance increase which is at least at large as the effect of using very large language models.

We obtained an f-score which outperforms other models reported in literature operating on identical data, even though we use vastly fewer features than others do.

Acknowledgements

This work was supported was supported under Australian Research Council's Discovery Projects funding scheme (project number DP110102593) and by the Australian Research Council as part of the Thinking Head Project the Thinking Head Project, ARC/NHMRC Special Research Initiative Grant # TS0669874. We thank the anonymous reviewers for their helpful comments.

References

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Version 1. Published by *Linguistic Data Consortium*, Philadelphia.

Erik Brill and Michele Banko. 2001. Mitigating the Paucity-of-Data Problem: Exploring the Effect of Training Corpus Size on Classifier Performance for Natural Language Processing. In *Proceedings of the First International Conference on Human Language Technology Research*.

Eugene Charniak and Mark Johnson. 2001. Edit detection and parsing for transcribed speech. In *Proceedings of the 2nd Meeting of the North American Chapter of the Association for Computational Linguistics*, pages 118–126.

Christopher Cieri David, David Miller, and Kevin Walker. 2004. Fisher English Training Speech Part 1 Transcripts. Published by *Linguistic Data Consortium*, Philadelphia.

Christopher Cieri David, David Miller, and Kevin Walker. 2005. Fisher English Training Speech Part 2 Transcripts. Published by *Linguistic Data Consortium*, Philadelphia.

John J. Godfrey and Edward Holliman. 1997. Switchboard-1 Release 2. Published by *Linguistic Data Consortium, Philadelphia*.

David Graff and Christopher Cieri. 2003. English gigaword. Published by *Linguistic Data Consortium*, Philadelphia.

Martin Jansche. 2005. Maximum Expected F-Measure Training of Logistic Regression Models. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 692–699, Vancouver, British Columbia, Canada, October. Association for Computational Linguistics.

Mark Johnson and Eugene Charniak. 2004. A TAG-based noisy channel model of speech repairs. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 33–39.

Mark Johnson, Eugene Charniak, and Matthew Lease. 2004. An Improved Model for Recognizing Disfluencies in Conversational Speech. In *Proceedings of the Rich Transcription Fall Workshop*.

Jeremy G. Kahn, Matthew Lease, Eugene Charniak, Mark Johnson, and Mari Ostendorf. 2005. Effective Use of Prosody in Parsing Conversational Speech. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 233–240, Vancouver, British Columbia, Canada.

Reinhard Kneser and Hermann Ney. 1995. Improved backing-off for m-gram language modeling. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, pages 181–184.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. 1999. Treebank-3. Published by *Linguistic Data Consortium, Philadelphia*.

William Schuler, Samir AbdelRahman, Tim Miller, and Lane Schwartz. 2010. Broad-Coverage Parsing using Human-Like Memory Constraints. *Computational Linguistics*, 36(1):1–30.

Richard Schwartz, Long Nguyen, Francis Kubala, George Chou, George Zavaliagkos, and John Makhoul. 1994. On Using Written Language Training Data for Spoken Language Modeling. In *Proceedings of the Human Language Technology Workshop*, pages 94–98.

Elizabeth Shriberg and Andreas Stolcke. 1998. How far do speakers back up in repairs? A quantitative model. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2183–2186.

Elizabeth Shriberg. 1994. *Preliminaries to a Theory of Speech Disuencies*. Ph.D. thesis, University of California, Berkeley.

David A. Smith and Jason Eisner. 2006. Minimum Risk Annealing for Training Log-Linear Models. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 787–794.

Matthew Snover, Bonnie Dorr, and Richard Schwartz. 2004. A Lexically-Driven Algorithm for Disfluency Detection. In *Proceedings of Human Language Technologies and North American Association for Computational Linguistics*, pages 157–160.

Andreas Stolcke. 2002. SRILM - An Extensible Language Modeling Toolkit. In *Proceedings of the International Conference on Spoken Language Processing*, pages 901–904.

Qi Zhang, Fuliang Weng, and Zhe Feng. 2006. A progressive feature selection algorithm for ultra large feature spaces. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 561–568.

ACL-HLT 2011

Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies

Volume 1: Long Papers

We wish to thank our sponsors

PLATINUM SPONSORS

research at Google

Baidu 百度
百度一下 你就知道

GOLD SPONSOR

Microsoft Research

SILVER SPONSORS

Pacific Northwest NATIONAL LABORATORY
Proudly Operated by Battelle *Since 1965*

YAHOO! LABS

BRONZE SPONSORS

Raytheon BBN Technologies

Appen ButlerHill

ETS Listening. Learning. Leading.®

W UNIVERSITY *of* WASHINGTON

IBM Research

OHSU OREGON HEALTH & SCIENCE UNIVERSITY
CENTER FOR SPOKEN LANGUAGE UNDERSTANDING

OHSU OREGON HEALTH & SCIENCE UNIVERSITY
DEPARTMENT OF MEDICAL INFORMATICS AND CLINICAL EPIDEMIOLOGY

SUPPORTERS

Rakuten

WIDMER BROTHERS

viaLanguage. say what you mean®

Preferred Infrastructure.

at&t

SDL

ACL HLT 2011

The 49th Annual Meeting of the
Association for Computational Linguistics:
Human Language Technologies

Proceedings of the Conference

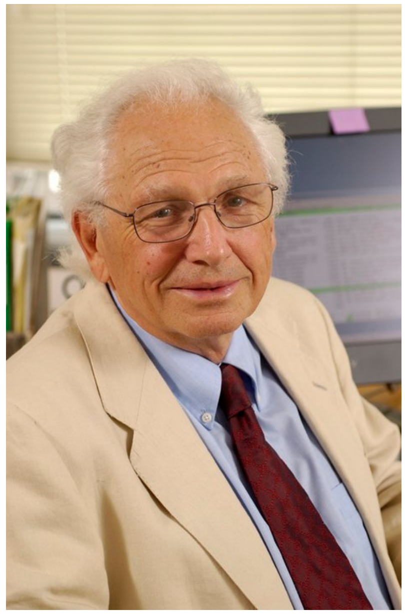
19-24 June, 2011
Portland, Oregon, USA

Production and Manufacturing by
Omnipress, Inc.
2600 Anderson Street
Madison, WI 53704 USA

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-932432-87-9



We dedicate the ACL 2011 proceedings to the memory of Fred Jelinek (1932-2010), who received ACL's Lifetime Achievement Award in 2009. His award acceptance speech can be found in Computational Linguistics 35(4), and an obituary by Mark Liberman appeared in Computational Linguistics 36(4). Several other newspaper and professional society obituaries have described his extraordinary personal life and career.

Fred's influence on computational linguistics is almost impossible to overstate. In the 1970s and 1980s, he and his colleagues at IBM developed the statistical paradigm that dominates our field today, including a great many specific techniques for modeling, parameter estimation, and search that continue to enjoy wide use. Even more fundamentally, as Mark Liberman recounts in his obituary, Fred led the field away from a mode where lone inventors defended their designs by appealing to aesthetics and anecdotes, to a more communal and transparent process of evaluating methods objectively through controlled comparisons on training and test sets.

Under Fred's visionary leadership, the IBM group revolutionized speech recognition by adopting a statistical, data-driven perspective that was deeply at odds with the rationalist ethos of the time. The group began with Fred's information-theoretic reconceptualization of the task as recovering a source signal (text) after it had passed through a noisy channel. They then worked out the many components needed for a full speech recognizer, along with the training algorithms for each component and global decoding algorithms. Steve Young, in an obituary in the IEEE SLTC Newsletter, describes Fred as not a pioneer but the pioneer of speech recognition.

In the 1980s, the IBM speech group's work on language modeling drew them toward deeper analysis of text. Fred and his colleagues introduced NLP methods such as word clustering, HMM part-of-speech tagging, history-based parsing, and prefix probability computation in PCFGs. They famously turned their noisy-channel lens on machine translation, founding the field of statistical MT with a series of ingenious and highly influential models.

After Fred moved to Johns Hopkins University in 1993, he worked tirelessly to improve language modeling by incorporating syntactic and other long-range dependencies as well as semantic classes. He also presided for 16 years over the Johns Hopkins Summer Workshops, whose 51 teams from 1995-2010 attacked a wide range of topics in human language technology, many making groundbreaking advances in the field.

There is a popular conception that Fred was somehow hostile to linguistics. Certainly he liked to entertain others by repeating his 1988 quip that "Any time a linguist leaves the group, the recognition rate goes up." Yet he had tried to leave information theory for linguistics as early as 1962, influenced by Noam Chomsky's lectures and his wife Milena's earlier studies with Roman Jakobson. He always strove for clean formal models just as linguists do. He was deeply welcoming toward any attempt to improve models through better linguistics, as long as they had a large number of parameters. Indeed, it was one of the major frustrations of his career that it was so difficult to beat n-gram language models, when humans were evidently using additional linguistic and world knowledge to obtain much better predictive performance. As he said in an award acceptance speech in 2004, "My colleagues and I always hoped that linguistics will eventually allow us to strike gold."

Fred was skeptical only about the relevance of armchair linguistics to engineering, believing that there was far more variation in the data than could be described compactly by humans. For this reason, while he was quite interested in recovering or exploiting latent linguistic structure, he trusted human-annotated linguistic data to be a better description of that structure than human-conceived linguistic rules. Statistical models could be aided even by imperfect or incomplete annotations, such as unaligned orthographic transcriptions, bilingual corpora, or syntactic analyses furnished by ordinary speakers. Fred pushed successfully for the development of such resources, notably the IBM/Lancaster Treebank and its successor, the Penn Treebank.

Fred influenced many of us personally. He was warm-hearted, witty, cultured, thoughtful about the scientific process, a generous mentor, and always frank, honest, and unpretentious. The changes that he brought to our field are largely responsible for its recent empirical progress and commercial success. They have also helped make it attractive to many bright, technically sophisticated young researchers. This proceedings volume, which is dedicated to his memory, testifies to the overwhelming success of his leadership and vision.

By Jason Eisner, on behalf of ACL 2011 Organizing Committee

Preface: General Chair

Welcome to the 49th Annual Meeting of the Association for Computational Linguistics in Portland, Oregon. ACL is perhaps the longest-running conference series in computer science. Amazingly, it is still growing. We expect this year's ACL to attract an even larger number of participants than usual, since 2011 happens to be an off-year for COLING, EACL and NAACL.

The yearly success of ACL results from the dedication and hard work of many people. This year is no exception. I would like to thank all of them for volunteering their time and energy in service to our community.

I thank the Program Co-Chairs Rada Mihalcea and Yuji Matsumoto for putting together a wonderful main conference program, including 164 long papers, 128 short papers and much anticipated keynote speeches by David Ferrucci and Lera Boroditsky. Tutorial Co-Chairs, Patrick Pantel and Andy Way solicited proposals and selected six fascinating tutorials in a wide range of topics. The Workshop Co-Chairs, Hal Daume III and John Carroll, organized a joint selection process with EMNLP 2011. The program consists of 3 two-day workshops and 13 one-day workshops, a new record number for ACL. Sadao Kurohashi, Chair of System Demonstrations, assembled a committee and oversaw the review of 46 demo submissions.

The Student Session is organized by Co-Chairs, Sasa Petrovic, Emily Pitler, Ethan Selfridge and Faculty Advisors: Miles Osborne, Thamar Solorio. They introduced a new, poster-only format to be held in conjunction with the main ACL poster session. They also obtained NSF funding to provide travel support for all student session authors.

Special thank goes to Publication Chair, Guodong Zhou and his assistant Hong Yu. They produced the entire proceedings of the conference.

We are indebted to Brain Roark and the local arrangement committee for undertaking a phenomenal amount detailed work over the course of two years to host this conference, such as allocating appropriate space to meet all the needs of the scientific program, compiling and printing of the conference handbook, arranging a live tango band for the banquet and dance, to name just a few. The local arrangement committee consists of: Nate Bodenstab (webmeister), Peter Heeman (exhibitions), Christian Monson (student volunteers), Zak Shafran and Meg Mitchell (social), Richard Sproat (local sponsorship), Mahsa Yarmohammadi and Masoud Rouhizadeh (student housing coordinators) and Aaron Dunlop (local publications coordinator).

I want to express my gratitude to Ido Dagan, Chair of the ACL Conference Coordination Committee, Dragomir Radev, ACL Secretary, and Priscilla Rasmussen, ACL Business Manager, for their advice and guidance throughout the process.

ACL 2011 has two Platinum Sponsors (Google and Baidu), one Gold Sponsor (Microsoft), two Silver sponsors (Pacific Northwest National Lab and Yahoo!), and seven Bronze Sponsors and six Supporters. We are grateful for the financial support from these organizations. I would like to thank and applaud the tremendous effort by the ACL sponsorship committee: Srinivas Bangalore (AT&T), Massimiliano Ciaramita (Google), Kevin Duh (NTT), Michael Gamon (Microsoft), Stephen Pulman (Oxford), Priscilla Rasmussen (ACL), and Haifeng Wang (Baidu).

Finally, I would like to thank all the area chairs, workshop organizers, tutorial presenters, authors, reviewers and conference attendees for their participation and contribution. I hope everyone will have a great time sharing ideas and inspiring one another at this conference.

ACL 2011 General Chair
Dekang Lin, Google, Inc.

Preface: Program Committee Co-Chairs

Welcome to the program of the 2011 Conference of the Association for Computational Linguistics! ACL continues to grow, and this year the number of paper submissions broke once again the record set by previous years. We received a total of 1,146 papers, out of which 634 were submitted as long papers and 512 were submitted as short papers. 25.7

To achieve the goal of a broad technical program, we followed the initiative from last year and solicited papers under four main different categories: *theoretical computational linguistics*, *empirical/data-driven approaches*, *resources/evaluation*, and *applications/tools*. We also continued to accept other types of papers (e.g., surveys or challenge papers), although unlike the previous year, no separate category was created for these papers. The papers falling under one of the four categories were reviewed using specialized reviewed forms; we also had a general review form that was used to review the papers that did not fall under one of the four main categories.

A new initiative this year was to also accept papers accompanied by supplemental materials (software and/or datasets). In addition to the regular review of the research quality of the paper, the accompanied resources were also reviewed for their quality, and the acceptance or rejection decisions were made based on the quality of both the paper and the supplemental materials. Among all the submissions, a total of 84 papers were accompanied by a software package and 117 papers were accompanied by a dataset. Among all the accepted papers, 30 papers are accompanied by software and 35 papers are accompanied by a dataset. These materials will be hosted on the ACL web site under http://www.aclweb.org/supplementals.

We are delighted to have two distinguished invited speakers: Dr. David Ferrucci (Principal Investigator, IBM Research), who will talk about his team's work on building *Watson* – a deep question answering system that achieved champion-level performance at Jeopardy!, and Lera Boroditsky (Assistant Professor, Stanford University), who will give a presentation on her research on how the languages we speak shape the way we think. In addition, the recipient of the ACL Lifetime Achievement Award will present a plenary lecture during the final day of the conference.

As in previous years, there will be three awards, one for the best long paper, one for the best long paper by a student, and one for the best short paper. The candidates for the best paper awards were nominated by the area chairs, who took into consideration the feedback they received from the reviewers on whether a paper might merit a best paper prize. From among the nominations we received, we selected the top five candidates for the long and short papers, and the final awards were then selected by the area chairs together with the program co-chairs. The recipients of the best paper awards will present their papers in a plenary session during the second day of the conference.

There are many individuals to thank for their contributions to the conference program. First and foremost, we would like to thank the authors who submitted their work to ACL. The growing number of submissions reflects how broad and active our field is. We are deeply indebted to the area chairs and the reviewers for their hard work. They enabled us to select an exciting program and to provide valuable feedback to the authors. We thank the general conference chair Dekang Lin and the local arrangements committee headed by Brian Roark for their help and advice, as well as last year's program committee co-chairs, Stephen Clark and Sandra Carberry, for sharing their experiences. Additional thanks go to the publications chair, Guodong Zhang, who put this volume together, and Yu Hong, who helped him with this task.

We are most grateful to Priscilla Rasmussen, who helped us with various logistic and organizational aspects of the conference. Rich Gerber and the START team responded to our questions quickly, and helped us manage the large number of submissions smoothly.

Enjoy the conference!

ACL 2011 Program Co-Chairs
Yuji Matsumoto, Nara Institute of Science and Technology
Rada Mihalcea, University of North Texas

Organizing Committee

General Chair

Dekang Lin, Google

Local Arrangements Chair

Brian Roark, Oregon Health & Science University

Program Co-Chairs

Yuji Matsumoto, Nara Institute of Science and Technology
Rada Mihalcea, University of North Texas

Local Arrangements Committee

Nate Bodenstab, Oregon Health & Science University
Aaron Dunlop, Oregon Health & Science University
Peter Heeman, Oregon Health & Science University
Meg Mitchell, Oregon Health & Science University
Christian Monson, Nuance
Zak Shafran, Oregon Health & Science University
Richard Sproat, Oregon Health & Science University
Masoud Rouhizadeh, Oregon Health & Science University
Mahsa Yarmohammadi, Oregon Health & Science University

Publications Chair

Guodong Zhou, Suzhou University

Sponsorship Chairs

Haifeng Wang, Baidu
Kevin Duh, National Inst. of Information and Communications Technology
Massimiliano Ciaramita, Google
Michael Gamon, Microsoft
Priscilla Rasmussen, Association for Computational Linguistics
Srinivas Bangalore, AT&T
Stephen Pulman, Oxford University

Tutorial Co-chairs

Patrick Pantel, Microsoft Research
Andy Way, Dublin City University

Workshop Co-chairs

Hal Daume III, University of Maryland
John Carroll, University of Sussex

Demo Chair

Sadao Kurohashi, Kyoto University

Mentoring

Chair
- Tim Baldwin, University of Melbourne

Committee
- Chris Biemann, TU Darmstadt
- Mark Dras, Macquarie University
- Jeremy Nicholson, University of Melbourne

Student Research Workshop

Student Co-chairs
- Sasa Petrovic, University of Edinburgh
- Emily Pitler, University of Pennsylvania
- Ethan Selfridge, Oregon Health & Science University

Faculty Advisors
- Miles Osborne, University of Edinburgh
- Thamar Solorio, University of Alabama at Birmingham

ACL Conference Coordination Committee

Ido Dagan, Bar Ilan University (chair)
Chris Brew, Ohio State University
Graeme Hirst, University of Toronto
Lori Levin, Carnegie Mellon University
Christopher Manning, Stanford University
Dragomir Radev, University of Michigan
Owen Rambow, Columbia University
Priscilla Rasmussen, Association for Computational Linguistics
Suzanne Stevenson, University of Toronto

ACL Business Manager

Priscilla Rasmussen, Association for Computational Linguistics

Program Committee

Program Co-chairs

Yuji Matsumoto, Nara Institute of Science and Technology
Rada Mihalcea, University of North Texas

Area Chairs

Razvan Bunescu, Ohio University
Xavier Carreras, Technical University of Catalonia
Anna Feldman, Montclair University
Pascale Fung, Hong Kong University of Science and Technology
Chu-Ren Huang, Hong Kong Polytechnic University
Kentaro Inui, Tohoku University
Greg Kondrak, University of Alberta
Shankar Kumar, Google
Yang Liu, University of Texas at Dallas
Bernardo Magnini, Fondazione Bruno Kessler
Elliott Macklovitch, Marque d'Or
Katja Markert, University of Leeds
Lluis Marquez, Technical University of Catalonia
Diana McCarthy, Lexical Computing Ltd
Ryan McDonald, Google
Alessandro Moschitti, University of Trento
Vivi Nastase, Heidelberg Institute for Theoretical Studies
Manabu Okumura, Tokyo Institute of Technology
Vasile Rus, University of Memphis
Fabrizio Sebastiani, National Research Council of Italy
Michel Simard, National Research Council of Canada
Thamar Solorio, University of Alabama at Birmingham
Svetlana Stoyanchev, Open University
Carlo Strapparava, Fondazione Bruno Kessler
Dan Tufis, Romanian Academy of Artificial Intelligence
Xiaojun Wan, Peking University
Taro Watanabe, National Inst. of Information and Communications Technology
Alexander Yates, Temple University
Deniz Yuret, Koc University

Program Committee

Ahmed Abbasi, Eugene Agichtein, Eneko Agirre, Lars Ahrenberg, Gregory Aist, Enrique Alfonseca, Laura Alonso i Alemany, Gianni Amati, Alina Andreevskaia, Ion Androutsopoulos, Abhishek Arun, Masayuki Asahara, Nicholas Asher, Giuseppe Attardi, Necip Fazil Ayan

Collin Baker, Jason Baldridge, Tim Baldwin, Krisztian Balog, Carmen Banea, Verginica Barbu

Mititelu, Marco Baroni, Regina Barzilay, Roberto Basili, John Bateman, Tilman Becker, Lee Becker, Beata Beigman-Klebanov, Cosmin Bejan, Ron Bekkerman, Daisuke Bekki, Kedar Bellare, Anja Belz, Sabine Bergler, Shane Bergsma, Raffaella Bernardi, Nicola Bertoldi, Pushpak Bhattacharyya, Archana Bhattarai, Tim Bickmore, Chris Biemann, Dan Bikel, Alexandra Birch, Maria Biryukov, Alan Black, Roi Blanco, John Blitzer, Phil Blunsom, Gemma Boleda, Francis Bond, Kalina Bontcheva, Johan Bos, Gosse Bouma, Kristy Boyer, S.R.K. Branavan, Thorsten Brants, Eric Breck, Ulf Brefeld, Chris Brew, Ted Briscoe, Samuel Brody

Michael Cafarella, Aoife Cahill, Chris Callison-Burch, Rafael Calvo, Nicoletta Calzolari, Nicola Cancedda, Claire Cardie, Giuseppe Carenini, Claudio Carpineto, Marine Carpuat, Xavier Carreras, John Carroll, Ben Carterette, Francisco Casacuberta, Helena Caseli, Julio Castillo, Mauro Cettolo, Hakan Ceylan, Joyce Chai, Pi-Chuan Chang, Vinay Chaudhri, Berlin Chen, Ying Chen, Hsin-Hsi Chen, John Chen, Colin Cherry, David Chiang, Yejin Choi, Jennifer Chu-Carroll, Grace Chung, Kenneth Church, Massimiliano Ciaramita, Philipp Cimiano, Stephen Clark, Shay Cohen, Trevor Cohn, Nigel Collier, Michael Collins, John Conroy, Paul Cook, Ann Copestake, Bonaventura Coppola, Fabrizio Costa, Koby Crammer, Dan Cristea, Montse Cuadros, Silviu-Petru Cucerzan, Aron Culotta, James Curran

Walter Daelemans, Robert Damper, Hoa Dang, Dipanjan Das, Hal Daume, Adria de Gispert, Marie-Catherine de Marneffe, Gerard de Melo, Maarten de Rijke, Vera Demberg, Steve DeNeefe, John DeNero, Pascal Denis, Ann Devitt, Giuseppe Di Fabbrizio, Mona Diab, Markus Dickinson, Mike Dillinger, Bill Dolan, Doug Downey, Markus Dreyer, Greg Druck, Kevin Duh, Chris Dyer, Marc Dymetman

Markus Egg, Koji Eguchi, Andreas Eisele, Jacob Eisenstein, Jason Eisner, Michael Elhadad, Tomaz Erjavec, Katrin Erk, Hugo Escalante, Andrea Esuli

Hui Fang, Alex Chengyu Fang, Benoit Favre, Anna Feldman, Christiane Fellbaum, Donghui Feng, Raquel Fernandez, Nicola Ferro, Katja Filippova, Jenny Finkel, Seeger Fisher, Margaret Fleck, Dan Flickinger, Corina Forascu, Kate Forbes-Riley, Mikel L. Forcada, Eric Fosler-Lussier, Jennifer Foster, George Foster, Anette Frank, Alex Fraser, Dayne Freitag, Guohong Fu, Hagen Fuerstenau, Pascale Fung, Sadaoki Furui

Evgeniy Gabrilovich, Robert Gaizauskas, Michel Galley, Michael Gamon, Kuzman Ganchev, Jianfeng Gao, Claire Gardent, Thomas Gärtner, Albert Gatt, Dmitriy Genzel, Kallirroi Georgila, Carlo Geraci, Pablo Gervas, Shlomo Geva, Daniel Gildea, Alastair Gill, Dan Gillick, Jesus Gimenez, Kevin Gimpel, Roxana Girju, Claudio Giuliano, Amir Globerson, Yoav Goldberg, Sharon Goldwater, Carlos Gomez Rodriguez, Julio Gonzalo, Brigitte Grau, Stephan Greene, Ralph Grishman, Tunga Gungor, Zhou GuoDong, Iryna Gurevych, David Guthrie

Nizar Habash, Ben Hachey, Barry Haddow, Gholamreza Haffari, Aria Haghighi, Udo Hahn, Jan Hajic, Dilek Hakkani-Tür, Keith Hall, Jirka Hana, John Hansen, Sanda Harabagiu, Mark Hasegawa-Johnson, Koiti Hasida, Ahmed Hassan, Katsuhiko Hayashi, Ben He, Xiaodong He, Ulrich Heid, Michael Heilman, Ilana Heintz, Jeff Heinz, John Henderson, James Henderson, Iris

Hendrickx, Aurelie Herbelot, Erhard Hinrichs, Tsutomu Hirao, Julia Hirschberg, Graeme Hirst, Julia Hockenmaier, Tracy Holloway King, Bo-June (Paul) Hsu, Xuanjing Huang, Liang Huang, Jimmy Huang, Jian Huang, Chu-Ren Huang, Juan Huerta, Rebecca Hwa

Nancy Ide, Gonzalo Iglesias, Gabriel Infante-López, Diana Inkpen, Radu Ion, Elena Irimia, Pierre Isabelle, Mitsuru Ishizuka, Aminul Islam, Abe Ittycheriah, Tomoharu Iwata

Martin Jansche, Sittichai Jiampojamarn, Jing Jiang, Valentin Jijkoun, Richard Johansson, Mark Johnson, Aravind Joshi

Nanda Kambhatla, Min-Yen Kan, Kyoko Kanzaki, Rohit Kate, Junichi Kazama, Bill Keller, Andre Kempe, Philipp Keohn, Fazel Keshtkar, Adam Kilgarriff, Jin-Dong Kim, Su Nam Kim, Brian Kingsbury, Katrin Kirchhoff, Ioannis Klapaftis, Dan Klein, Alexandre Klementiev, Kevin Knight, Rob Koeling, Oskar Kohonen, Alexander Kolcz, Alexander Koller, Kazunori Komatani, Terry Koo, Moshe Koppel, Valia Kordoni, Anna Korhonen, Andras Kornai, Zornitsa Kozareva, Lun-Wei Ku, Sandra Kuebler, Marco Kuhlmann, Roland Kuhn, Mikko Kurimo, Oren Kurland, Olivia Kwong

Krista Lagus, Philippe Langlais, Guy Lapalme, Mirella Lapata, Dominique Laurent, Alberto Lavelli, Matthew Lease, Gary Lee, Kiyong Lee, Els Lefever, Alessandro Lenci, James Lester, Gina-Anne Levow, Tao Li, Shoushan LI, Fangtao Li, Zhifei Li, Haizhou Li, Hang Li, Wenjie Li, Percy Liang, Chin-Yew Lin, Frank Lin, Mihai Lintean, Ken Litkowski, Diane Litman, Marina Litvak, Yang Liu, Bing Liu, Qun Liu, Jingjing Liu, Elena Lloret, Birte Loenneker-Rodman, Adam Lopez, Annie Louis, Xiaofei Lu, Yue Lu

Tengfei Ma, Wolfgang Macherey, Klaus Macherey, Elliott Macklovitch, Nitin Madnani, Bernardo Magnini, Suresh Manandhar, Gideon Mann, Chris Manning, Daniel Marcu, David Martínez, Andre Martins, Yuval Marton, Sameer Maskey, Spyros Matsoukas, Mausam, Arne Mauser, Jon May, David McAllester, Andrew McCallum, David McClosky, Ryan McDonald, Bridget McInnes, Tara McIntosh, Kathleen McKeown, Paul McNamee, Yashar Mehdad, Qiaozhu Mei, Arul Menezes, Paola Merlo, Donald Metzler, Adam Meyers, Haitao Mi, Jeff Mielke, Einat Minkov, Yusuke Miyao, Dunja Mladenic, Marie-Francine Moens, Saif Mohammad, Dan Moldovan, Diego Molla, Christian Monson, Manuel Montes y Gomez, Raymond Mooney, Robert Moore, Tatsunori Mori, Glyn Morrill, Sara Morrissey, Alessandro Moschitti, Jack Mostow, Smaranda Muresan, Gabriel Murray, Gabriele Musillo, Sung-Hyon Myaeng

Tetsuji Nakagawa, Mikio Nakano, Preslav Nakov, Ramesh Nallapati, Vivi Nastase, Borja Navarro-Colorado, Roberto Navigli, Mark-Jan Nederhof, Matteo Negri, Ani Nenkova, Graham Neubig, Guenter Neumann, Vincent Ng, Hwee Tou Ng, Patrick Nguyen, Jian-Yun Nie, Rodney Nielsen, Joakim Nivre, Tadashi Nomoto, Scott Nowson

Diarmuid Ó Séaghdha, Sharon O'Brien, Franz Och, Stephan Oepen, Kemal Oflazer, Jong-Hoon Oh, Constantin Orasan, Miles Osborne, Gozde Ozbal

Sebastian Pado, Tim Paek, Bo Pang, Patrick Pantel, Soo-Min Pantel, Ivandre Paraboni, Cecile Paris, Marius Pasca, Gabriella Pasi, Andrea Passerini, Rebecca J. Passonneau, Siddharth Patwardhan, Adam Pauls, Adam Pease, Ted Pedersen, Anselmo Penas, Anselmo Peñas, Jing Peng, Fuchun Peng, Gerald Penn, Marco Pennacchiotti, Wim Peters, Slav Petrov, Emanuele Pianta, Michael Picheny, Daniele Pighin, Manfred Pinkal, David Pinto, Stelios Piperidis, Paul Piwek, Benjamin Piwowarski, Massimo Poesio, Livia Polanyi, Simone Paolo Ponzetto, Hoifung Poon, Ana-Maria Popescu, Andrei Popescu-Belis, Maja Popovic, Martin Potthast, Richard Power, Sameer Pradhan, John Prager, Rashmi Prasad, Partha Pratim Talukdar, Adam Przepiórkowski, Vasin Punyakanok, Matthew Purver, Sampo Pyysalo

Silvia Quarteroni, Ariadna Quattoni, Chris Quirk

Stephan Raaijmakers, Dragomir Radev, Filip Radlinski, Bhuvana Ramabhadran, Ganesh Ramakrishnan, Owen Rambow, Aarne Ranta, Delip Rao, Ari Rappoport, Lev Ratinov, Antoine Raux, Emmanuel Rayner, Roi Reichart, Ehud Reiter, Steve Renals, Philip Resnik, Giuseppe Riccardi, Sebastian Riedel, Stefan Riezler, German Rigau, Ellen Riloff, Laura Rimell, Eric Ringger, Horacio Rodríguez, Paolo Rosso, Antti-Veikko Rosti, Rachel Edita Roxas, Alex Rudnicky, Marta Ruiz Costa-Jussa, Vasile Rus, Graham Russell, Anton Rytting

Rune Sætre, Kenji Sagae, Horacio Saggion, Tapio Salakoski, Agnes Sandor, Sudeshna Sarkar, Anoop Sarkar, Giorgio Satta, Hassan Sawaf, Frank Schilder, Anne Schiller, David Schlangen, Sabine Schulte im Walde, Tanja Schultz, Holger Schwenk, Donia Scott, Yohei Seki, Satoshi Sekine, Stephanie Seneff, Jean Senellart, Violeta Seretan, Burr Settles, Serge Sharoff, Dou Shen, Wade Shen, Libin Shen, Kiyoaki Shirai, Luo Si, Grigori Sidorov, Mário Silva, Fabrizio Silvestri, Khalil Simaan, Michel Simard, Gabriel Skantze, Noah Smith, Matthew Snover, Rion Snow, Benjamin Snyder, Stephen Soderland, Marina Sokolova, Thamar Solorio, Swapna Somasundaran, Lucia Specia, Valentin Spitkovsky, Richard Sproat, Manfred Stede, Mark Steedman, Amanda Stent, Mark Stevenson, Svetlana Stoyanchev, Veselin Stoyanov, Michael Strube, Sara Stymne, Keh-Yih Su, Fangzhong Su, Jian Su, L Venkata Subramaniam, David Suendermann, Maosong Sun, Mihai Surdeanu, Richard Sutcliffe, Charles Sutton, Jun Suzuki, Stan Szpakowicz, Idan Szpektor

Hiroya Takamura, David Talbot, Irina Temnikova, Michael Tepper, Simone Teufel, Stefan Thater, Allan Third, Jörg Tiedemann, Christoph Tillmann, Ivan Titov, Takenobu Tokunaga, Kentaro Torisawa, Kristina Toutanova, Isabel Trancoso, Richard Tsai, Vivian Tsang, Dan Tufis

Takehito Utsuro

Shivakumar Vaithyanathan, Alessandro Valitutti, Antal van den Bosch, Hans van Halteren, Gertjan van Noord, Lucy Vanderwende, Vasudeva Varma, Tony Veale, Olga Vechtomova, Paola Velardi, Rene Venegas, Ashish Venugopal, Jose Luis Vicedo, Evelyne Viegas, David Vilar, Begona Villada Moiron, Sami Virpioja, Andreas Vlachos, Stephan Vogel, Piek Vossen

Michael Walsh, Xiaojun Wan, Xinglong Wang, Wei Wang, Haifeng Wang, Justin Washtell, Andy

Way, David Weir, Ben Wellner, Ji-Rong Wen, Chris Wendt, Michael White, Ryen White, Richard Wicentowski, Jan Wiebe, Sandra Williams, Jason Williams, Theresa Wilson, Shuly Wintner, Kam-Fai Wong, Fei Wu

Deyi Xiong, Peng Xu, Jinxi Xu, Nianwen Xue

Scott Wen-tau Yih, Emine Yilmaz

David Zajic, Fabio Zanzotto, Richard Zens, Torsten Zesch, Hao Zhang, Bing Zhang, Min Zhang, Huarui Zhang, Jun Zhao, Bing Zhao, Jing Zheng, Li Hai Zhou, Michael Zock, Andreas Zollmann, Geoffrey Zweig, Pierre Zweigenbaum

Secondary Reviewers

Omri Abend, Rodrigo Agerri, Paolo Annesi, Wilker Aziz, Tyler Baldwin, Verginica Barbu Mititelu, David Batista, Delphine Bernhard, Stephen Boxwell, Janez Brank, Chris Brockett, Tim Buckwalter, Wang Bukang, Alicia Burga, Steven Burrows, Silvia Calegari, Marie Candito, Marina Cardenas, Bob Carpenter, Paula Carvalho, Diego Ceccarelli, Asli Celikyilmaz, Soumaya Chaffar, Bin Chen, Danilo Croce, Daniel Dahlmeier, Hong-Jie Dai, Mariam Daoud, Steven DeNeefe, Leon Derczynski, Elina Desypri, Sobha Lalitha Devi, Gideon Dror, Loic Dugast, Eraldo Fernandes, Jody Foo, Kotaro Funakoshi, Jing Gao, Wei Gao, Diman Ghazi, Julius Goth, Joseph Grafsgaard, Eun Young Ha, Robbie Haertel, Matthias Hagen, Enrique Henestroza, Hieu Hoang, Maria Holmqvist, Dennis Hoppe, Yunhua Hu, Yun Huang, Radu Ion, Elena Irimia, Jagadeesh Jagarlamudi, Antonio Juárez-González, Sun Jun, Evangelos Kanoulas, Aaron Kaplan, Caroline Lavecchia, Lianhau Lee, Michael Levit, Ping Li, Thomas Lin, Wang Ling, Ying Liu, José David Lopes, Bin Lu, Jia Lu, Saab Mansour, Raquel Martinez-Unanue, Haitao Mi, Simon Mille, Teruhisa Misu, Behrang Mohit, Sílvio Moreira, Rutu Mulkar-Mehta, Jason Naradowsky, Sudip Naskar, Heung-Seon Oh, You Ouyang, Lluís Padró, Sujith Ravi, Marta Recasens, Luz Rello, Stefan Rigo, Alan Ritter, Alvaro Rodrigo, Hasim Sak, Kevin Seppi, Aliaksei Severyn, Chao Shen, Shuming Shi, Laurianne Sitbon, Jun Sun, György Szarvas, Eric Tang, Alberto Téllez-Valero, Luong Minh Thang, Gabriele Tolomei, David Tomás, Diana Trandabat, Zhaopeng Tu, Gokhan Tur, Kateryna Tymoshenko, Fabienne Venant, Esaú Villatoro-Tello, Joachim Wagner, Dan Walker, Wei Wei, Xinyan Xiao, Jun Xie, Hao Xiong, Gu Xu, Jun Xu, Huichao Xue, Taras Zagibalov, Beñat Zapirain, Kalliopi Zervanou, Renxian Zhang, Daqi Zheng, Arkaitz Zubiaga

Table of Contents

Conference Program

Monday, June 20, 2011

(8:45-9:00) Opening Session

(9:00-10:00) Invited Talk 1:IBM Watson Deep QA System (tentative title) by David Ferrucci

(10:00-10:30) Coffee Break

Session 1-A: (10:30-12:10) MT: Methods

A Word-Class Approach to Labeling PSCFG Rules for Machine Translation
Andreas Zollmann and Stephan Vogel

Deciphering Foreign Language
Sujith Ravi and Kevin Knight

Effective Use of Function Words for Rule Generalization in Forest-Based Translation
Xianchao Wu, Takuya Matsuzaki and Jun'ichi Tsujii

Combining Morpheme-based Machine Translation with Post-processing Morpheme Prediction
Ann Clifton and Anoop Sarkar

Session 1-B: (10:30-12:10) Machine Learning Methods 1

Evaluating the Impact of Coder Errors on Active Learning
Ines Rehbein and Josef Ruppenhofer

A Fast and Accurate Method for Approximate String Search
Ziqi Wang, Gu Xu, Hang Li and Ming Zhang

Domain Adaptation by Constraining Inter-Domain Variability of Latent Feature Representation
Ivan Titov

Exact Decoding of Syntactic Translation Models through Lagrangian Relaxation
Alexander M. Rush and Michael Collins

Session 1-C: (10:30-12:10) Information Retrieval

Session 1-D: (10:30-12:10) Sentiment Analysis/Opinion Mining 1

Monday, June 20, 2011 (continued)

Session 1-E: (10:30-12:10) Language Resource

(12:10 - 2:00) Lunch

Session 2-A: (2:00-3:40) MT: Models & Evaluation

Monday, June 20, 2011 (continued)

Monday, June 20, 2011 (continued)

Monday, June 20, 2011 (continued)

(3:40-4:10) Coffee Break

Session 3-A: (4:10-5:50) Transliteration/Alignment

Session 3-B: (4:10-5:50) Parsing 1

Session 3-C: (4:10-5:50) Summarization

Session 3-D: (4:10-5:50) Relation Extraction

Monday, June 20, 2011 (continued)

Session 3-E: (4:10-5:50) Semantics

(6:00-8:30) Poster Session (Long papers)

(6:00-8:30) Poster Session (Short papers)

Tuesday, June 21, 2011

Session 4-A: (9:00-10:30) Best Paper Session

(10:30-11:00) Coffee Break

(3:30-4:00) Coffee Break

Session 7-A: (4:00-5:40) SMT: Phrase-based Models

Session 7-B: (4:00-5:40) Parsing 2

Tuesday, June 21, 2011 (continued)

Tuesday, June 21, 2011 (continued)

Session 7-E: (4:00-5:40) Coreference Resolution

Bootstrapping coreference resolution using word associations
Hamidreza Kobdani, Hinrich Schuetze, Michael Schiehlen and Hans Kamp

Large-Scale Cross-Document Coreference Using Distributed Inference and Hierarchical Models
Sameer Singh, Amarnag Subramanya, Fernando Pereira and Andrew McCallum

A Cross-Lingual ILP Solution to Zero Anaphora Resolution
Ryu Iida and Massimo Poesio

Coreference Resolution with World Knowledge
Altaf Rahman and Vincent Ng

(7:00-11:00) Banquet

Wednesday, June 22, 2011

(9:00-10:00) Invited Talk 2: How do the languages we speak shape the ways we think? by Lera Boroditsky

(10:00-10:30) Coffee Break

Session 5-A: (10:30-12:10) SMT: Tree-based Models

How to train your multi bottom-up tree transducer
Andreas Maletti

Binarized Forest to String Translation
Hao Zhang, Licheng Fang, Peng Xu and Xiaoyun Wu

Learning to Transform and Select Elementary Trees for Improved Syntax-based Machine Translations
Bing Zhao, Young-Suk Lee, Xiaoqiang Luo and Liu Li

Rule Markov Models for Fast Tree-to-String Translation
Ashish Vaswani, Haitao Mi, Liang Huang and David Chiang

Session 5-B: (10:30-12:10) Morphology/POS Induction

Session 5-C: (10:30-12:10) Error Correction

Wednesday, June 22, 2011 (continued)

Session 5-D: (10:30-12:10) Information Extraction

A Generative Entity-Mention Model for Linking Entities with Knowledge Base
Xianpei Han and Le Sun

Simple supervised document geolocation with geodesic grids
Benjamin Wing and Jason Baldridge

Piggyback: Using Search Engines for Robust Cross-Domain Named Entity Recognition
Stefan Rüd, Massimiliano Ciaramita, Jens Müller and Hinrich Schütze

Template-Based Information Extraction without the Templates
Nathanael Chambers and Dan Jurafsky

Session 5-E: (10:30-12:10) Discourse

Classifying arguments by scheme
Vanessa Wei Feng and Graeme Hirst

Automatically Evaluating Text Coherence Using Discourse Relations
Ziheng Lin, Hwee Tou Ng and Min-Yen Kan

Underspecifying and Predicting Voice for Surface Realisation Ranking
Sina Zarrieß, Aoife Cahill and Jonas Kuhn

Recognizing Authority in Dialogue with an Integer Linear Programming Constrained Model
Elijah Mayfield and Carolyn Penstein Rosé

Wednesday, June 22, 2011 (continued)

(12:10 - 2:00) Lunch

(1:30-3:00) ACL Business Meeting

(3:00-3:30) Coffee Break

Session 6-A: (3:30-4:45) MT: Reordering Models

Reordering Metrics for MT
Alexandra Birch and Miles Osborne

Reordering with Source Language Collocations
Zhanyi Liu, Haifeng Wang, Hua Wu, Ting Liu and Sheng Li

A Joint Sequence Translation Model with Integrated Reordering
Nadir Durrani, Helmut Schmid and Alexander Fraser

Session 6-B: (3:30-4:45) Grammar

Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results
Roger Levy

Metagrammar engineering: Towards systematic exploration of implemented grammars
Antske Fokkens

Simple Unsupervised Grammar Induction from Raw Text with Cascaded Finite State Models
Elias Ponvert, Jason Baldridge and Katrin Erk

Wednesday, June 22, 2011 (continued)

(5:00-6:10) Life time achievement award and closing

Monday, June 20, 2011

(6:00-8:30) Poster Session (Long papers)

Monday, June 20, 2011 (continued)

Minimum Bayes-risk System Combination
Jesús González-Rubio, Alfons Juan and Francisco Casacuberta

Adjoining Tree-to-String Translation
Yang Liu, Qun Liu and Yajuan Lü

Enhancing Language Models in Statistical Machine Translation with Backward N-grams and Mutual Information Triggers
Deyi Xiong, Min Zhang and Haizhou Li

Translating from Morphologically Complex Languages: A Paraphrase-Based Approach
Preslav Nakov and Hwee Tou Ng

Gappy Phrasal Alignment By Agreement
Mohit Bansal, Chris Quirk and Robert Moore

Translationese and Its Dialects
Moshe Koppel and Noam Ordan

Rare Word Translation Extraction from Aligned Comparable Documents
Emmanuel Prochasson and Pascale Fung

Using Bilingual Parallel Corpora for Cross-Lingual Textual Entailment
Yashar Mehdad, Matteo Negri and Marcello Federico

Using Large Monolingual and Bilingual Corpora to Improve Coordination Disambiguation
Shane Bergsma, David Yarowsky and Kenneth Church

Unsupervised Decomposition of a Document into Authorial Components
Moshe Koppel, Navot Akiva, Idan Dershowitz and Nachum Dershowitz

Discovering Sociolinguistic Associations with Structured Sparsity
Jacob Eisenstein, Noah A. Smith and Eric P. Xing

Local and Global Algorithms for Disambiguation to Wikipedia
Lev Ratinov, Dan Roth, Doug Downey and Mike Anderson

Monday, June 20, 2011 (continued)

Monday, June 20, 2011 (continued)

Monday, June 20, 2011 (continued)

Invited Talk 1: Building Watson: An Overview of the DeepQA Project

David Ferrucci, Principal Investigator, IBM Research
Monday, June 20 9:00-10:00

Computer systems that can directly and accurately answer peoples' questions over a broad domain of human knowledge have been envisioned by scientists and writers since the advent of computers themselves. Open domain question answering holds tremendous promise for facilitating informed decision making over vast volumes of natural language content. Applications in business intelligence, healthcare, customer support, enterprise knowledge management, social computing, science and government could all benefit from computer systems capable of deeper language understanding. The DeepQA project is aimed at exploring how advancing and integrating Natural Language Processing (NLP), Information Retrieval (IR), Machine Learning (ML), Knowledge Representation and Reasoning (KR&R) and massively parallel computation can greatly advance the science and application of automatic Question Answering. An exciting proof-point in this challenge was developing a computer system that could successfully compete against top human players at the Jeopardy! quiz show (www.jeopardy.com).

Attaining champion-level performance at Jeopardy! requires a computer to rapidly and accurately answer rich open-domain questions, and to predict its own performance on any given question. The system must deliver high degrees of precision and confidence over a very broad range of knowledge and natural language content with a 3-second response time. To do this, the DeepQA team advanced a broad array of NLP techniques to find, generate, evidence and analyze many competing hypotheses over large volumes of natural language content to build Watson (www.ibmwatson.com). An important contributor to Watson's success is its ability to automatically learn and combine accurate confidences across a wide array of algorithms and over different dimensions of evidence. Watson produced accurate confidences to know when to "buzz in" against its competitors and how much to bet. High precision and accurate confidence computations are critical for real business settings where helping users focus on the right content sooner and with greater confidence can make all the difference. The need for speed and high precision demands a massively parallel computing platform capable of generating, evaluating and combing 1000's of hypotheses and their associated evidence. In this talk, I will introduce the audience to the Jeopardy! Challenge, explain how Watson was built on DeepQA to ultimately defeat the two most celebrated human Jeopardy Champions of all time and I will discuss applications of the Watson technology beyond in areas such as healthcare.

Dr. David Ferrucci is the lead researcher and Principal Investigator (PI) for the Watson/Jeopardy! project. He has been a Research Staff Member at IBM's T.J. Watson's Research Center since 1995 where he heads up the Semantic Analysis and Integration department. Dr. Ferrucci focuses on technologies for automatically discovering valuable knowledge in natural language content and using it to enable better decision making.

Invited Talk 2: How do the languages we speak shape the ways we think?

Lera Boroditsky, Assistant Professor, Stanford University
Wednesday, June 22 9:00-10:00

Do people who speak different languages think differently? Does learning new languages change the way you think? Do polyglots think differently when speaking different languages? Are some thoughts unthinkable without language? I will describe data from experiments conducted around the world that reveal the powerful and often surprising ways that the languages we speak shape the ways we think.

Lera Boroditsky is an assistant professor of psychology at Stanford University and Editor in Chief of Frontiers in Cultural Psychology. Boroditsky's research centers on how knowledge emerges out of the interactions of mind, world, and language, and the ways that languages and cultures shape human thinking. To this end, Boroditsky's laboratory has collected data around the world, from Indonesia to Chile to Turkey to Aboriginal Australia. Her research has been widely featured in the media and has won multiple awards, including the CAREER award from the National Science Foundation, the Searle Scholars award, and the McDonnell Scholars award.