Multimodal Speaker Identification Based on Text and Speech

Panagiotis Moschonas and Constantine Kotropoulos

Department of Informatics, Aristotle University of Thessaloniki,
Box 451, Thessaloniki 54124, Greece
pmoschon@csd.auth.gr,costas@aiia.csd.auth.gr

Abstract. This paper proposes a novel method for speaker identification based on both speech utterances and their transcribed text. The transcribed text of each speaker's utterance is processed by the probabilistic latent semantic indexing (PLSI) that offers a powerful means to model each speaker's vocabulary employing a number of hidden topics, which are closely related to his/her identity, function, or expertise. Mel-frequency cepstral coefficients (MFCCs) are extracted from each speech frame and their dynamic range is quantized to a number of predefined bins in order to compute MFCC local histograms for each speech utterance, which is time-aligned with the transcribed text. Two identity scores are independently computed by the PLSI applied to the text and the nearest neighbor classifier applied to the local MFCC histograms. It is demonstrated that a convex combination of the two scores is more accurate than the individual scores on speaker identification experiments conducted on broadcast news of the RT-03 MDE Training Data Text and Annotations corpus distributed by the Linguistic Data Consortium.

Key words: multimodal speaker identification, text, speech, probabilistic latent semantic indexing, Mel-frequency cepstral coefficients, nearest neighbor classifier, convex combination

1 Introduction

Speaker identification systems resort mainly to speech processing. Undoubtedly, speech is probably the most natural modality to identify a speaker [1]. Historically in speaker recognition technology R&D, effort has been devoted to characterizing the statistics of a speaker's amplitude spectrum. Although, dynamic information (e.g., difference spectra) has been taken into consideration as well as static information, the focus has been on spectral rather than temporal characterization. The usage of certain words and phrases [2] as well as intonation, stress, and timing [3], constitute longer term speech patterns, which define "familiar-speaker" differences, a promising but radical departure from mainstream speaker recognition technology.

In this paper, we explore text that is rarely combined with speech for biometric person identification. More specifically, text refers to the time-aligned transcribed speech that appears as rich annotation of speakers' utterances. The annotation process could be an automatic, a semi-automatic, or a manual task as is frequently the case. In the proposed algorithm, we assume that we know the start time and end time of each word in a speaker's utterance as well as its speech to text transcription. Although there are a few past works where text was exploited for speaker identification, e.g. the idiolectal differences as quantified by N-gram language models [2], to the best of authors' knowledge no multimodal approach that exploits speech and text has been proposed so far.

The motivation for building multimodal biometric systems is that systems based on a single-modality, e.g. speech, are far from being error-free, especially under noisy operating conditions. The use of complementary modalities, such as visual speech, speaker's face, yields a more reliable identification accuracy. However, the additional modalities may also be unstable due to dependence on recording conditions, such as changes in pose and lighting conditions. Text and language models, if available, do not suffer from such shortcomings.

The transcribed text of each speaker's utterance is processed by the probabilistic latent semantic indexing (PLSI)[4] that offers a powerful means to model each speaker's vocabulary employing a number of hidden topics, which are closely related to his/her identity, function, or expertise. Mel-frequency cepstral coefficients (MFCCs) are extracted from each speech frame and their dynamic range is quantized to a number of predefined bins in order to compute MFCC local histograms for each speech utterance, that is time-aligned with the transcribed text. Two identity scores are independently computed by the PLSI applied first to the text and the nearest neighbor classifier applied next to the local MFCC histograms. It is demonstrated that a late fusion of the two scores by a convex combination is more accurate than the individual scores on closed-set speaker identification experiments conducted on broadcast news of the RT-03 MDE Training Data Text and Annotations corpus distributed by the Linguistic Data Consortium [6].

The outline of the paper is as follows. In Section 2, a novel method to combine audio and text data in a single representation array is described. Speaker identification algorithms based on either text or speech are described in Section 3. Experimental results are demonstrated in Section 4, and conclusions are drawn in Section 5.

2 Biometric Data Representation

In this Section, we propose a novel representation of speaker biometric data that will be used as an input to the identification algorithms to be described in the next section. As far as text data are concerned, two sets are identified, namely the set of speaker identities and the domain vocabulary. The latter is the union of all vocabularies used by the speakers. A closed set of speaker identities S of cardinality n is assumed, i.e.

$$S = \{s_1, s_2, \ldots, s_n\} . \tag{1}$$

Let W be the domain vocabulary of cardinality m:

$$W = \{w_1, w_2, \ldots, w_m\} \,. \tag{2}$$

A two dimensional matrix $\mathbf{K}$ whose rows refer to spoken words in W and its columns refer to the speaker identities in S is created. Its (i, j)-th element, $k_{i,j}$, is equal to the number of times the word w_i is uttered by the speaker s_j:

$$\mathbf{K} = \begin{bmatrix} k_{1,1} & k_{1,2} & \ldots & k_{1,n} \\ k_{2,1} & k_{2,2} & \ldots & k_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ k_{m,1} & k_{m,2} & \ldots & k_{m,n} \end{bmatrix} \,. \tag{3}$$

It is obvious that the “word-by-speaker” matrix $\mathbf{K}$ plays the same role with the “term-by-document” matrix in PLSI. The only difference is that the columns are associated to speakers and not to documents. Such a representation can be modeled in terms to latent variables, which refer to topics. The models can easily be derived by applying PLSI to $\mathbf{K}$. To minimize the vocabulary size, one may apply stemming or some sort of word clustering. Function words (e.g. articles, propositions) are frequently rejected as well.

Next, time-aligned audio information is associated with each element of the “word-by-speaker” matrix. This is done by extracting the MFCCs [5] for each frame within the speech utterance of each spoken word. Since, the same word might have been spoken by the same speaker more than once, we should aggregate the MFCC information from multiple instances of the same word. This is done as follows.

1. For each frame within each word utterance, extract 13 MFCCs. That is, 13 MFCC sets of variable length are obtained depending on the duration of each word utterance.
2. Create the histogram of each MFCC by splitting its dynamic range into b bins. Since we do not know a priori the dynamic range of each MFCC, we need to determine the minimum and maximum value for each MFCC.
3. Finally, add the MFCC histograms for all word utterances spoken by each speaker.

Accordingly, we obtain a $13 \times b$ matrix, where b is the number of histogram bins. Let the maximum and minimum value of each MFCC be $\max_c$ and $\min_c$,respectively, $c = 1, 2, \ldots, 13$. The size of each bin δb_c is given by

$$\delta b_c = \frac{\max_c - \min_c}{b}, \; c = 1, 2, \ldots, 13. \tag{4}$$

Let

$$\mathbf{A}_{i,j} = \begin{bmatrix} \alpha_{1,1;i,j} & \alpha_{1,2;i,j} & \cdots & \alpha_{1,b;i,j} \\ \alpha_{2,1;i,j} & \alpha_{2,2;i,j} & \cdots & \alpha_{2,b;i,j} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha_{13,1;i,j} & \alpha_{13,2;i,j} & \cdots & \alpha_{13,b;i,j} \end{bmatrix} \tag{5}$$

be the $13 \times b$ matrix whose element $\alpha_{c,t;i,j}$ denotes how many times the c-th MFCC is fallen into the t-th bin of the histogram for the i-th word spoken by the j-th speaker. It is proposed each element of the "word-by-speaker" matrix to index the pair $(k_{i,j}, \mathbf{A}_{i,j})$. If $k_{i,j} = 0$, then $\mathbf{A}_{i,j} = \mathbf{0}$. Consequently, Eq. (3) is rewritten as

$$\mathbf{K} = \begin{bmatrix} (k_{1,1}, \mathbf{A}_{1,1}) & (k_{1,2}, \mathbf{A}_{1,2}) & \dots & (k_{1,n}, \mathbf{A}_{1,n}) \\ (k_{2,1}, \mathbf{A}_{2,1}) & (k_{2,2}, \mathbf{A}_{2,2}) & \dots & (k_{2,n}, \mathbf{A}_{2,n}) \\ \vdots & \vdots & \ddots & \vdots \\ (k_{m,1}, \mathbf{A}_{m,1}) & (k_{m,2}, \mathbf{A}_{m,2}) & \dots & (k_{m,n}, \mathbf{A}_{m,n}) \end{bmatrix} . \tag{6}$$

The main advantage of the proposed multimodal biometric representation is that it can easily be updated when new data arrive. When a new word or a new speaker is added (e.g. during training), one has to add a new row or column in **K**, respectively. Another main characteristic of the data representation is that contains only integers. This has a positive impact in data storage, since in most cases, an unsigned integer needs 32 bits, whereas a double number needs 64bits [6].

3 Multimodal Speaker Identification

Having defined the biometric data representation, let us assume that the training data form the composite matrix **K** as in Eq. (6). Let the test data contain instances of speech and text information from a speaker $s_x \in S$ whose identity is to be determined. The test data are represented by the following composite vector $\mathbf{k}_x$, i.e.

$$\mathbf{k}_x = \begin{bmatrix} (k_{1,x}, \mathbf{A}_{1,x}) \\ (k_{2,x}, \mathbf{A}_{2,x}) \\ \vdots \\ (k_{m,x}, \mathbf{A}_{m,x}) \end{bmatrix} . \tag{7}$$

The composite matrix **K** and the composite vector $\mathbf{k}_x$ must have the same number of rows, thus the domain vocabulary should be the same. By denoting the vocabulary that is used by the test speaker as W_x, we could use the union of both training and test vocabulary:

$$W_{all} = W \cup W_x \; . \tag{8}$$

Accordingly, new rows might be inserted to both **K** and $\mathbf{k}_x$ and be rearranged so that each row is associated to the same word in the domain vocabulary. The next step is to combine the training and test data in one matrix as follows:

$$\mathbf{K}_{all} = \left[\mathbf{K} \mid \mathbf{k}_x \right] \; . \tag{9}$$

Having gathered all the data in the unified structure, $\mathbf{K}_{all}$, first PLSI is applied to its $k_{i,j}$ entries in order to reveal a distribution of topics related to the textual content uttered by each speaker in S. In the following, the topics are defined by the latent discrete random variable z that admits q values in the set

$$Z = \{z_1, z_2, \ldots, z_q\} \tag{10}$$

as in [4]. Let us denote by $P(s,z)$ the joint probability that speaker s speaks about topic z. Obviously,

$$P(s,z) = P(s|z)\, P(z), \tag{11}$$

where $P(s|z)$ is the conditional probability of a speaker given a topic and $P(z)$ is the probability of topic. By applying the PLSI algorithm, one can estimate the constituents of Eq. (11). The expectation step of the Expectation-Maximization algorithm (EM) in PLSI yields

$$P(z|w,s) = \frac{P(z)\, P(w|z)\, P(s|z)}{\sum\limits_{z'} P(z')\, P(w|z')\, P(s|z')}. \tag{12}$$

The maximization step is described by the following set of equations:

$$P(w|z) = \frac{\sum\limits_{s} k_{w,s}\, P(z|w,s)}{\sum\limits_{w',s} k_{w',s}\, P(z|w',s)} \tag{13}$$

$$P(s|z) = \frac{\sum\limits_{w} k_{w,s}\, P(z|w,s)}{\sum\limits_{w,s'} k_{w,s'}\, P(z|w,s')} \tag{14}$$

$$P(z) = \frac{1}{R} \sum_{w,s} k_{w,s}\, P(z|w,s) \tag{15}$$

where $R \equiv \sum\limits_{w,s} k_{w,s}$. The number of iterations of the EM algorithm can be preset by the user or can be determined by monitoring a convergence criterion, such as to observe insignificant changes of the model probabilities of PLSI. A random initialization of the model probabilities is frequently applied. The number of topics is also predetermined by the user.

Let the joint probability speaker $s_j \in S$ from the training set speaks about topic z_t be

$$P_{j,t} = P(s_j, z_t), \;\; 1 \leq j \leq n, \;\; 1 \leq t \leq q. \tag{16}$$

Similarly, let $P_{x,t} = P(s_x, z_t)$ be the same joint probability for the test speaker s_x. Then, we can define a distance between the speakers s_x and s_j based on text information as

$$d_{PLSI}(x,j) = \frac{1}{q} \sum_{t=1}^{q} |P_{j,t} - P_{x,t}| \quad j = 1, 2, \ldots, n \tag{17}$$

or

$$d_{PLSI}(x,j) = \frac{1}{q} \sum_{t=1}^{q} P_{j,t} \log \frac{P_{j,t}}{P_{x,t}} \quad j = 1, 2, \ldots, n. \tag{18}$$

Eq. (17) defines an L_1-norm, whereas Eq. (18) is the KullbackLeibler divergence of the joint probabilities of speakers and topics. By applying either distance, we can obtain a vector containing all distances between the test speaker s_x and all speakers $s_j \in S$:

$$\mathbf{D}_{PLSI}(x) = \left[d_{PLSI}(x,1)\; d_{PLSI}(x,2)\; \ldots\; d_{PLSI}(x,n)\right]^T \tag{19}$$

Let us now consider the definition of distances between speakers when local histograms of MFCCs are employed. First, we create the set of word indices L_j for each column of $\mathbf{K}$ (i.e., the training set):

$$L_j = \{i \mid k_{i,j} > 0\}\,,\; j = 1, 2, \ldots, n. \tag{20}$$

Similarly, let $L_x = \{i \mid k_{i,x} > 0\}$. A distance function between the local MFCC histograms stored in $\mathbf{A}_{i,x}$ and $\mathbf{A}_{i,j}$ can be defined as

$$d_{MFCC}(x,j) = \frac{1}{|L_x \cup L_j|} \sum_{i \in \left(L_x \cup L_j\right)} \left(\frac{1}{13b} \sum_{c_1=1}^{13} \sum_{c_2=1}^{b} |\alpha_{c_1,c_2;i,x} - \alpha_{c_1,c_2;i,j}| \right) \tag{21}$$

where $|L_x \cup L_j|$ is the number of common words used by speakers s_x and s_j, b denotes the chosen number of MFCC local histogram bins, and $\alpha_{c_1,c_2;i,j}$ refers to the c_2-th bin in the local histogram of the c_1-th MFCC at the i-th word spoken by the j-th speaker column. A vector $\mathbf{D}_{MFCC}(x)$ containing the distances between the test speaker s_x and all training speakers can be defined:

$$\mathbf{D}_{MFCC}(x) = \left[d_{MFCC}(x,1)\; d_{MFCC}(x,2)\; \ldots\; d_{MFCC}(x,n)\right]^T\ . \tag{22}$$

The elements of the distance vector in Eq. (22) can be normalized by dividing with the maximum value admitted by the distances. A convex combination of the distance vectors can be used to combine Eq. (19) and Eq. (22):

$$\mathbf{D}(x) = \gamma\, \mathbf{D}_{PLSI}(x) + (1-\gamma)\, \mathbf{D}_{MFCC}(x) \tag{23}$$

where the parameter $\gamma \in [0,1]$ weighs our confidence for the text-derived distance. As $\gamma \rightarrow 0$, the identification depends more on the information extracted from speech, whereas for $\gamma \rightarrow 1$ emphasis is given to the information extracted from text.

The algorithm ends by finding the minimum element value in $\mathbf{D}(x)$, whose index refers to the speaker that best matches s_x and accordingly it is assigned to s_x, i.e.:

$$s_x = \arg\min_j \left[\gamma\, d_{PLSI}(x,j) + (1-\gamma)\, d_{MFCC}(x,j)\right]\ . \tag{24}$$

4 Experimental Results

To demonstrate the proposed multimodal speaker identification algorithm, experiments are conducted on broadcast news (BN) collected within the DARPA

Efficient, Affordable, Reusable Speech-to-Text (EARS) Program in Metadata Extraction (MDE). That is, a subset of the so called RT-03 MDE Training Data Text and Annotations corpus [7] is used. BN enable to easily assess the algorithm performance, because each speaker has a specific set of topics to talk about. The BN speech data were drawn from the 1997 English Broadcast News Speech (HUB4) corpus. HUB4 stem from four distinct sources, namely the American Broadcasting Company, the National Broadcasting Company, Public Radio International and the Cable News Network. Overall, the transcripts and annotations cover approximately 20 hours of BN. In the experiments conducted, the total duration of the speech recordings exceeds 2 hours.

To begin with, let us first argue on the motivation for combining text and speech. Figs. 1 and 2 demonstrate two cases for 14 and 44 speakers, where the average speaker identification rate increases by combining PLSI applied to text and nearest neighbor classifier applied to MFCC histograms.

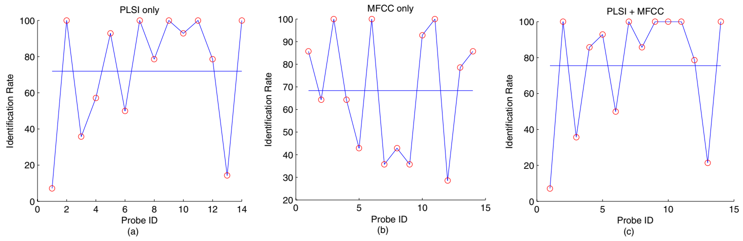

Fig. 1. Identification rate versus Probe ID when 14 speakers are employed. Average identification rates for (a) PLSI: 72%; (b) MFCCs: 68%; (c) Both: 75%.

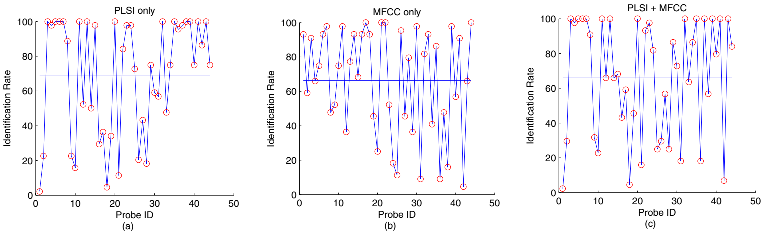

Fig. 2. Identification rate versus Probe ID when 44 speakers are employed. Average identification rates for (a) PLSI: 69%; (b) MFCCs: 66%; (c) Both: 67%.

Next, two sets of experiments are conducted. Both sets contain three experiments with a varying number of speakers. Speech and text modalities are treated equally. That is, $\gamma = 0.5$ in Eq. (24). Fig. 3 shows the percentage of correctly identified speakers within the R best matches for $R = 1, 2, \ldots, 20$, i.e., the so called cumulative match score versus rank curve after having performed 100 iterations and chosen 4 latent topics in PLSI as well as 10 bins for each MFCC histogram. As we can see, the algorithm produces near perfect identification for 20 speakers. Concerning the group of the 37 speakers, the results are satisfactory after the 4th rank. The more difficult case, when identification among 90 speakers is sought, reveals a poor, but acceptable performance, especially after the 7th rank.

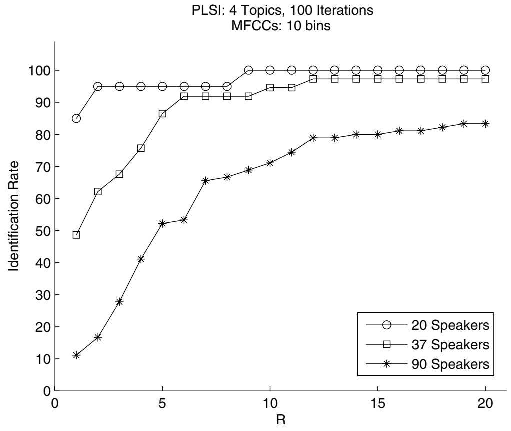

Fig. 3. Cumulative match score versus rank curve of the proposed algorithm using 4 topics and 100 iterations in PLSI model and 10 bins for every MFCC histogram.

For comparison purposes, the percentage of correctly identified speakers within the R best matches using only PLSI for the same number of iterations and topics is plotted in Figure 4. The multimodal identification offers self-evident gains for best match identification in the case of small and medium sized speaker sets, while slight improvements of 3.32% are measured for the large speaker set.

In the second set of experiments, the proposed identification algorithm is fine tuned by increasing the number of iterations to 250, the number of topics to 12, and the number of histogram bins to 50. Although, such an increase has a negative impact on the speed of the algorithm, the results are improved considerably in some cases. From the comparison of Figures 3 and 5 it is seen that the identification rate for 20 speakers is slightly increased for the best match. For the medium-sized group of 37 speakers, the identification rate for the best match is climbed at nearly 70% from 50% in the previous set. For the large group of 90 speakers, the identification rate for the best match remains the same.

By repeating the identification using only PLSI with 12 topics and 250 iterations, the percentage of correctly identified speakers within the R best matches shown in Figure 6 is obtained. The comparison of Figures 5 and 6 validates

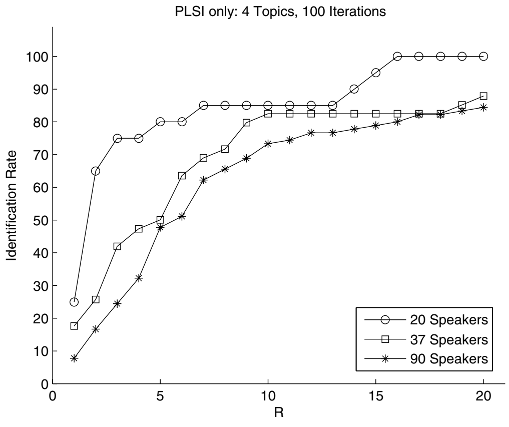

Fig. 4. Cumulative match score versus rank curve of PLSI using 4 topics and 100 iterations.

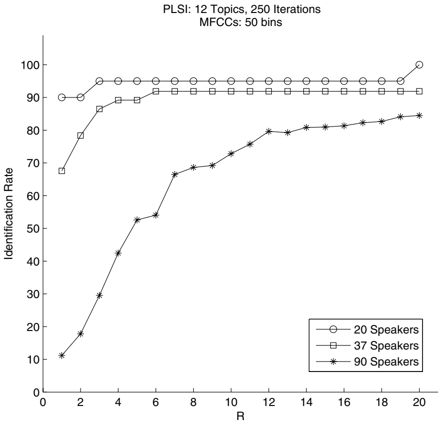

Fig. 5. Cumulative match score versus rank curve of the proposed algorithm using 12 topics and 250 iterations in PLSI model and 50 bins for every MFCC histogram.

that the identification rate at best match using both text and speech increases considerably for small and medium sized speaker sets, while marginal gains are obtained for large speaker sets. Moreover, the increased number of latent topics and iterations in PLSI have helped PLSI to improve its identification rate.

5 Conclusions

In this paper, first promising speaker identification rates have been reported by combining in a late fusion scheme text-based and speech-based distances in

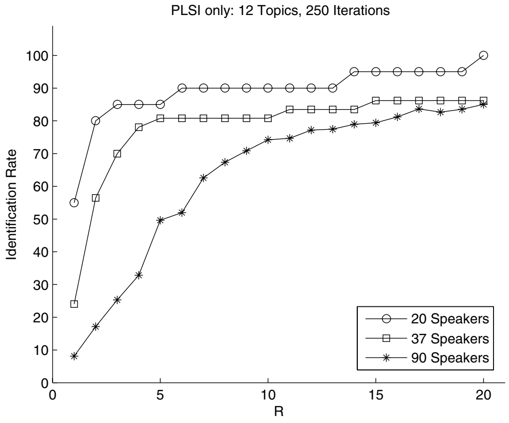

Fig. 6. Cumulative match score versus rank curve of PLSI only using 12 topics and 250 iterations.

experiments conducted on broadcast news of the RT-03 MDE Training Data Text and Annotations corpus. Motivated by the promising results, we plan to integrate MFCC histograms and document word histograms in PLSI, since both features are of the same nature and to study their early fusion.

Acknowledgments. This work has been performed within the COST Action 2101 on Biometrics for Identity Documents and Smart Cards.

References

1. Campbell, P. J.: "Speaker recognition: A turorial," *Proceedings of the IEEE*, vol. 85, no. 9, pp. 1437-1462 (1997).
2. Doddington, G: "Speaker recognition based on idiolectal differences between speakers," In Proc. *Eurospeech*, pp. 2521-2524 (2001).
3. Weber, F., Manganaro, L., Peskin, B., and Shriberg, E.: "Using prosodic and lexical information for speaker identification," In Proc. *2002 IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, vol. I, pp. 141-144 (2002).
4. Hofmann, T.: "Probabilistic latent semantic indexing.," In: Proc. *22nd Annual Int. Conf. Research and Development in Information Retrieval (SIGIR-99)*, pp. 50-57 (1999).
5. Davies, S. B. and Mermelstein, P.: "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions Acoustic, Speech, and Signal Processing*, vol. 31, pp. 793-807, (1983).
6. Schildt, H.: *C++ The Complete Reference*, 4/e. Osborne/McGraw-Hill, N. Y. (2002).
7. Strassel, S., Walker, C., and Lee H.: RT-03 MDE Training Data Text and Annotations, Linguistic Data Consortium (LDC), Philadelphia (2004).