PARAPHRASTIC LANGUAGE MODELS AND COMBINATION WITH NEURAL NETWORK LANGUAGE MODELS

X. Liu, M. J. F. Gales & P. C. Woodland

Cambridge University Engineering Dept,
Trumpington St., Cambridge, CB2 1PZ U.K.
Email: {*xl207,mjfg,pcw*}*@eng.cam.ac.uk*

ABSTRACT

In natural languages multiple word sequences can represent the same underlying meaning. Only modelling the observed surface word sequence can result in poor context coverage, for example, when using n-gram language models (LM). To handle this issue, paraphrastic LMs were proposed in previous research and successfully applied to a US English conversational telephone speech transcription task. In order to exploit the complementary characteristics of paraphrastic LMs and neural network LMs (NNLM), the combination between the two is investigated in this paper. To investigate paraphrastic LMs' generalization ability to other languages, experiments are conducted on a Mandarin Chinese broadcast speech transcription task. Using a paraphrastic multi-level LM modelling both word and phrase sequences, significant error rate reductions of 0.9% absolute (9% relative) and 0.5% absolute (5% relative) were obtained over the baseline n-gram and NNLM systems respectively, after a combination with word and phrase level NNLMs.

Index Terms: language model, paraphrase, speech recognition

1. INTRODUCTION

Natural languages have layered structures, a deeper structure that represents the meaning and core semantic relations of a sentence, and a surface form found in normal written texts or speech. The mapping from the meaning to surface form involves a natural language generation process and is often one-to-many. Multiple surface word sequences can be used to convey identical or similar semantic information. They are paraphrastic to each other, but use different syntactic, lexical and morphological rules in generation. Only modelling the observed surface word sequence can result in poor context coverage, for example, when using n-gram language models (LM).

To handle this problem, it is possible to directly model paraphrase variants when constructing the LM. Since alternative expressions of the same meaning are considered, the resulting LM's context coverage and generalization performance is expected to be improved. Along this line, the use of word level synonym features [8, 10, 7, 3] has been investigated. However, there are two issues associated with these existing approaches. First, the paraphrastic relationship between longer span syntactic structures, such as phrases, is largely ignored. Hence, a more general form of modelling that can also capture a higher level paraphrase mapping is preferred. Second, previous research focused on using manually derived expert semantic labelling provided by resources such as WordNet [5]. As manual annotation is usually very expensive, the scope of applying these methods to large tasks or rare resource languages is limited. Automatic, statistical paraphrase induction and extraction techniques are thus required. In order to address these issues, a novel form of language model, the paraphrastic LM, was proposed in [15], and successfully applied to a state-of-the-art LVCSR task for US English conversational telephone speech.

Both paraphrastic LMs and neural network LMs [23] can be used to improve LM generalization. However, there are major differences between them that can also be exploited as complementary characteristics. First, paraphrastic LMs can be trained using a large amounts of training data. In contrast, to reduce computational cost, NNLMs are normally trained using only a small in-domain data set, for example, audio transcripts. Secondly, paraphrastic LMs redistribute sufficient statistics to variable length paraphrase variants of the same sentence. The resulting sequence level smoothing of LM probabilities is different from the n-gram level smoothing used by NNLMs. Finally, the paraphrastic LMs considered in this paper are based on n-gram models. Despite being more efficient than NNLMs in probability computation, their generalization ability remain limited for unseen contexts that can not be found in either the training data or the associated paraphrases. A back-off to lower order distributions is still required. Techniques that can represent n-gram probabilities in a continuous space, such as NNLMs, can alleviate this problem [23, 20]. Hence, in order to to leverage the strengths of both models, the combination between paraphrastic LMs and NNLMs is investigated in this paper. In order to further investigate paraphrastic LMs' generalization ability to other languages, experiments are conducted on a Mandarin Chinese broadcast speech transcription task.

The rest of the paper is organized as follows. Paraphrastic LMs are introduced in section 2. The paraphrase extraction and lattice generation schemes are reviewed in section 3. The combination between paraphrastic LMs and neural network LMs is proposed in section 4. In section 5 a range of paraphrastic LMs and their combination with NNLMs are evaluated on a state-of-the-art Mandarin Chinese broadcast speech transcription task. The main contributions of this paper, relationship to previous work in the field and possible future work are summarized in section 6.

2. PARAPHRASTIC LANGUAGE MODELS

As discussed in section 1, in order to capture the paraphrastic relationship between longer span syntactic structures, a more general form of modelling should be used. To address this issue, the particular type of LMs proposed in this paper can flexibly model paraphrase mapping at the word, phrase and sentence level. As LM probabilities are estimated in the paraphrased domain, they are referred to as *paraphrastic language models* (PLM) [15]. For a L word long sentence

The research leading to these results was supported by EPSRC Programme Grant EP/I031022/1 (Natural Speech Technology).

$\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$ in the training data, the marginal probability over all paraphrase variant sequences is maximized,

$$\mathcal{F}(\mathcal{W}) = \ln \left(\sum_{\boldsymbol{\psi}, \boldsymbol{\psi}', \mathcal{W}'} P(\mathcal{W}|\boldsymbol{\psi}) P(\boldsymbol{\psi}|\boldsymbol{\psi}') P(\boldsymbol{\psi}'|\mathcal{W}') P_{\mathsf{PLM}}(\mathcal{W}') \right) \quad (1)$$

where

- $P_{\mathsf{PLM}}(\mathcal{W}')$ is paraphrastic LM probability to be estimated.
- $P(\boldsymbol{\psi}'|\mathcal{W}')$ is a word to phrase segmentation model assigning the probability of a phrase level segmentation, $\boldsymbol{\psi}'$, given a paraphrase word sequence $\mathcal{W}'$;
- $P(\boldsymbol{\psi}|\boldsymbol{\psi}')$ is a phrase to phrase paraphrase model computing the probability of a phrase sequence $\boldsymbol{\psi}$ being paraphrastic to another $\boldsymbol{\psi}'$;
- $P(\mathcal{W}|\boldsymbol{\psi})$ is a phrase to word segmentation model that converts a phrase sequence $\boldsymbol{\psi}$ to a word sequence $\mathcal{W}$, and by definition is a deterministic, one-to-one mapping, thus considered non-informative.

It can be shown that the sufficient statistics for a maximum likelihood (ML) estimation of $P_{\mathsf{PLM}}(\mathcal{W}')$ are accumulated along each paraphrase word sequence and weighted by its posterior probability. For a particular n-gram predicting word w_i following history h_i, the associated statistics $C(h_i, w_i)$ are

$$C(h_i, w_i) \;=\; \sum_{\mathcal{W}'} P(\mathcal{W}'|\mathcal{W}) C_{\mathcal{W}'}(h_i, w_i) \quad (2)$$

where $C_{\mathcal{W}'}(h_i, w_i)$ is the count of subsequence $<h_i, w_i>$ occurring in paraphrase variant $\mathcal{W}'$. During word to phrase segmentation, ambiguity can occur. If there is no clear reason to favor one phrase segmentation over another, $P(\boldsymbol{\psi}'|\mathcal{W}')$ may be treated as non-informative, as is considered in this work.

As sufficient statistics are discounted and re-distributed to alternative expressions of the same word sequence, paraphrastic LMs are expected to have a richer context coverage and broader distribution, but at the same time potentially increased modelling confusion than conventional LMs trained on the surface word sequence. One approach to balance the specific, but lower coverage word-based N-gram LMs with a more generic LM is to linearly interpolate the LM probabilities. This is commonly used with class-based LMs [18] and is used in this paper with paraphrastic LMs. Let $P(\tilde{w}|\tilde{h})$ denote the interpolated LM probability for any in-vocabulary word $\tilde{w}$ following an arbitrary history $\tilde{h}$, this is given by

$$P(\tilde{w}|\tilde{h}) \;=\; \lambda_{\mathsf{NG}} P_{\mathsf{NG}}(\tilde{w}|\tilde{h}) + \lambda_{\mathsf{PLM}} P_{\mathsf{PLM}}(\tilde{w}|\tilde{h}) \quad (3)$$

where λ_{NG} and λ_{PLM} are the interpolation weights assigned to the conventional LM distribution $P_{\mathsf{NG}}(\cdot)$ and the paraphrastic LM $P_{\mathsf{PLM}}(\cdot)$. They can be optimized on some held-out data.

In order to increase the context span for paraphrastic LMs, a phrase level paraphrastic LM can also be trained. This can be obtained by optimizing a simplified form of the criterion in equation (1), by dropping the word to phrase segmentation model $P(\boldsymbol{\psi}'|\mathcal{W}')$,

$$\mathcal{F}(\mathcal{W}) = \ln \left(\sum_{\boldsymbol{\psi}, \boldsymbol{\psi}'} P(\mathcal{W}|\boldsymbol{\psi}) P(\boldsymbol{\psi}|\boldsymbol{\psi}') P_{\mathsf{PLM}}(\boldsymbol{\psi}') \right) \quad (4)$$

thus the sufficient statistics in equation (2) accumulated on phrase level instead. In order to incorporate richer linguistic constraints, it is possible to train and log-linearly combine LMs that model different units, for example, words and phrases. LMs built at word and phrase level are log-linearly combined to yield a multi-level LM to further improve discrimination [12, 14]. This requires word level lattices to be first converted to phrase level lattices before the log-linear combination is performed. The log-linear interpolation weights were set equal for word and phrase level LMs, and kept fixed for all experiments of this paper.

3. PARAPHRASE PHRASE PAIR EXTRACTION AND PARAPHRASE LATTICE GENERATION

As discussed in sections 1 and 2, a phrase level paraphrase model is used in paraphrastic LMs. In order to obtain sufficient phrase coverage, an appropriate technique to learn a large number of paraphrase phrase pairs is required. Since it is impractical to obtain expert semantic labelling at the phrase, or sentence level as for SMT tasks [22], statistical paraphrase extraction schemes are needed [1, 16]. Techniques that perform paraphrase pair extraction from standard text data [11, 21] can be used. These are motivated by the *distributional similarity* theory [6], which postulates that phrase pairs often sharing the same left and right contexts are likely to be paraphrases to each other. As standard text data in large amounts can be used, wide phrase coverage can be obtained. Due to this advantage, the n-gram paraphrase induction algorithm presented in [15] is used to estimate the paraphrase model. The same minimum and maximum phrase length and the left and right context length settings were also used for all experiments in this paper. This algorithm can be extended to incorporate additional useful information, for example, syntactic constraints. In common with other paraphrase induction methods, the above scheme can also produce phrase pairs that are non-paraphrastic, for example, antonyms. However, this is of less concern for language modelling, for which improving context coverage is the prime aim.

In order to train paraphrastic LMs, multiple paraphrase variants are required to compute the sufficient statistics given in equation (2), as discussed in section 2. As all four components of the paraphrastic LM given in equation (1) can be efficiently represented by weighted finite state transducers (WFST) [17], rather than designing special purpose decoding tools, the WFST based decoding approach proposed in [15] was used in paraphrase lattice generation. The statistics required for paraphrastic LM estimation are then accumulated from the paraphrase lattices via a forward-backward pass. Using the WFST based decoding approach and a paraphrase model trained on 545 million words of English conversational data, for an example sentence some paraphrase variants generated are shown below.

Original Sentence:			
AND	I	GENERALLY	PREFER
Paraphrases:			
AND	I	REALLY	LIKE
I MEAN	I		WOULD LIKE
I GUESS	I	GENERALLY	LIKE
YOU KNOW	I	JUST	WANT
SO	I		APPRECIATE
I THINK	I		NEED
‘CAUSE	I		LOVE
WELL	I		PREFER
UM	I		WISH
...	...	...	...
Antonyms:			
AND YOU KNOW	I		HATE

As the paraphrase extraction method can also produce phrase pairs that are non-paraphrastic, antonym word sequences such as "AND YOU KNOW I HATE" were also found in the paraphrase lattice, as is shown in the bottom of the table.

In order to improve phrase coverage, expert semantic labelling provided by resources such as WordNet [5], and HowNet [4] for the Chinese language, when available, can also be used to generate paraphrases. As these paraphrase phrase pairs are not statistically derived, the resulting paraphrase model are treated as non-informative. Due to their different nature, statistically learned and expert derived paraphrase pairs were used to generate separate sets of lattices, and paraphrastic LMs. These models are then used in the interpolation with standard LMs in equation (3), as considered in this work.

4. COMBINING PARAPHRASTIC LANGUAGE MODELS WITH NEURAL NETWORK LANGUAGE MODELS

As discussed in section 1, paraphrastic LMs differ significantly from neural network LMs [23] in terms of the training data used, model structure and probability estimation. These differences can also be exploited as complementary characteristics. Hence, it is possible to appropriately combine paraphrastic LMs with NNLMs to leverage the strengths of both models. The particular form of combination considered in this paper is a linear interpolation between the paraphrastic LM, the NNLM and the conventional n-gram LM. The interpolated LM probabilities given in equation (3) is modified as,

$$P(\tilde{w}|\tilde{h}) = \lambda_{\mathsf{NG}} P_{\mathsf{NG}}(\tilde{w}|\tilde{h}) + \lambda_{\mathsf{PLM}} P_{\mathsf{PLM}}(\tilde{w}|\tilde{h}) + \lambda_{\mathsf{NN}} P_{\mathsf{NN}}(\tilde{w}|\tilde{h}) \quad (5)$$

where λ_{NN} is the interpolation weight assigned to the neural network LM. In the same fashion as in equation (3), component LM interpolation weights can be optimized on held-out data.

For the multi-level paraphrastic LMs discussed in section 2, the above interpolation needs to be performed at both word and phrase level prior to the log-linear combination between the word and phrase level LMs. In addition to a word level neural network LM, a neural network LM constructed using phrase level segmented training data is also required.

To reduce computational cost, conventional NNLMs only model the probabilities of a more frequently occurring subset of the complete vocabulary, commonly referred to as the *shortlist* [23]. The output layer normally only contains nodes for in-shortlist words. A similar approach may also be used at the input layer. Two issues arise when using this conventional NNLM architecture. First, NNLM parameters are trained only using the statistics of in-shortlist words thus introduces an undue bias to them. Secondly, as there is no explicit modelling of probabilities of *out-of-shortlist* (OOS) words in the output layer, statistics associated with them are also discarded in NNLM training. To handle these issues, an NNLM architecture with an additional output node explicitly modelling the probability mass of OOS words [19] is used in this paper. This ensures that all training data are used in NNLM training, and the probabilities of in-shortlist words are smoothed by the OOS probability mass, thus obtaining a more robust parameter estimation.

5. EXPERIMENTS AND RESULTS

In this section performance of various paraphrastic language models are evaluated on the CU-HTK LVCSR system for Mandarin Chinese broadcast speech used in the 2011 DARPA GALE evaluation. The system was trained on 1960 hours of broadcast speech data released by the LDC. A 63k recognition word list was used in decoding. The system uses a multi-pass recognition framework. In the initial lattice generation stage, adapted gender dependent cross-word triphone MPE acoustic models with HLDA projected PLP features augmented with pitch features, and an interpolated 3-gram word level baseline LM were used. A detailed description of the baseline system can be found in [13]. A 3 hour GALE Chinese speech test set, **dev09s**, of mixed broadcast news (BN) and conversation (BC) genres was used. For all results presented in this paper, matched pairs sentence-segment word error (MAPSSWE) based statistical significance test was performed at a significance level $\alpha = 0.05$.

The baseline LM was trained using a total of 5.9 billion characters from 28 difference text sources. These account for 4 billion words after a longest first based character to word segmentation was applied. The four text sources with the highest interpolation weights, the GALE acoustic transcriptions, **BN** (0.13) and **BC** (0.31), of 20 million words in total, the LDC GigaWord Xinhua News data, **GigaXin** (0.16), of 680 million words, and the Gale web data **GaleWeb** of 800 million words (0.09), were used to build various language models. These LMs are then evaluated to measure the character error rate (CER) via lattice rescoring. The 4 billion word full set trained 4-gram LM gave an error rate of 10.3% on **dev09s**, while a comparable 4-gram LM trained using only the above four text sources produced a competitive CER score of 10.4% and was used as a baseline in the following experiments, as is shown in the 1st line of table 2. Using this baseline system the BN and BC genre specific performance on **d09s** are 5.4% and 15.2% respectively.

Information on corpus size, paraphrase extraction schemes used and the number of phrase pairs extracted from the these four text sources, as well as HowNet [4], an expert semantic database for the Chinese language, are given in table 1. Using the automatic n-gram paraphrase extraction scheme discussed in section 3, a total of 80k, 35.9M and 1.8M phrase pairs were extracted from the **BN+BC**, **GigaXin** and **GaleWeb** data respectively. The expert semantic labelling by HowNet gave a total of 3.7M paraphrase phrase pairs.

Source	Size	Extraction	# Phrase Pairs
HowNet	-	Expert	3.7M
BN+BC	20M	Automatic	80k
GigaXin	680M	Automatic	35.9M
GaleWeb	800M	Automatic	1.8M

Table 1. Text size, paraphrase extraction method and the number of phrase pairs extracted from different Chinese text data sources.

The word level paraphrastic 4-gram LM, as is shown in the 4th line of table 2, outperformed the word level baseline LM "w4g" (shown in 1st line of table 2) by 0.3%, and the class LM baseline (shown in 2nd line of table 2) 0.2% absolute respectively. The two multi-level LMs in table 2 both used a total of 503k distinct multi-word phrases found in the paraphrase phrase table in addition to the baseline 63k word list. The baseline non-paraphrastic multi-level LM, shown in 3rd line of table 2, was trained on the phrase level text data obtained using a longest available word to phrase segmentation. The paraphrastic multi-level LM, shown in the last line of table 2, outperformed its comparable non-paraphrastic baseline by 0.2% absolute. Using this paraphrastic multi-level LM, an overall significant CER reduction of 0.5% absolute was obtained over the word level 4-gram baseline LM. It also outperformed the word level 4-gram baseline LM trained using the full training set of 4 billion words, with 24 more text sources, by 0.4% absolute. These results, together with those previously reported in [15] for English conversational telephone speech, confirm the cross-genre generalization ability of

paraphrastic LMs, as well as their scalability when large amounts of data is used in training.

LM	Paraphrastic	dev09s
w4g	×	10.4
w4g+clslm		10.3
w4g ∘ p4g		10.1
w4g	√	10.1
w4g ∘ p4g		9.9

Table 2. Performance of LMs trained using **BN**, **BC**, **GigaXin** and **GaleWeb** data on **dev09s**. "w4g" denotes word level 4-gram LM, "w4g+clslm" a word level 4-gram LM interpolated with a class LM with 1000 classes, and "w4g ∘ p4g" a multi-level LM log-linearly combining word and phrase level 4-gram LMs.

So far in this paper, paraphrastic LMs have been used to improve the performance of n-gram LMs. As discussed in section 4, it's also interesting to investigate the combination between paraphrastic LMs and state-of-the-art language modelling techniques, such as neural network LMs [23]. A total of four LMs shown in table 2, including the baseline 4-gram word level LM, its paraphrastic counterpart, and the two multi-level LMs, were combined with various neural network LMs using the method presented in section 4. A word level 4-gram NNLM with an OOS output layer node [19] was trained using the 20 million words of the **BN+BC** acoustic transcription only. The size of the NNLM input and output vocabularies are 45k and 20k words respectively. The associated coverage rate of its input and output vocabularies on the test data are 100% and 97%. A phrase level 4-gram NNLM was also trained using the same data, and the same phrase segmentation used by the multi-level LMs of table 2. Its input and output vocabularies contain 100k and 20k most frequent phrases. The coverage rate of its input and output vocabularies on the phrase level segmented test data are 98% and 85% respectively. For both the word and phrase level NNLMs, a total of 600 projection layer nodes and 400 hidden layer nodes were used. The performance of the baseline and the paraphrastic 4-gram word level LMs without any interpolation with NNLMs, are shown in the 1st and 4th lines of table 3 (also previously shown in the 1st and 4th lines of table 2).

LM	Paraphrastic n-gram LM	dev09s
w4g	×	10.4
w4g+$\mathrm{nn_w}$		10.0
(w4g+$\mathrm{nn_w}$) ∘ (p4g+$\mathrm{nn_p}$)		9.8
w4g	√	10.1
w4g+$\mathrm{nn_w}$		9.7
(w4g+$\mathrm{nn_w}$) ∘ (p4g+$\mathrm{nn_p}$)		9.5

Table 3. Performance of LMs trained using **BN**, **BC**, **GigaXin** and **GaleWeb** data on **dev09s**. "w4g" denotes word level 4-gram LM, "w4g+$\mathrm{nn_w}$" a word level 4-gram LM interpolated with an NNLM, and "(w4g+$\mathrm{nn_w}$) ∘ (p4g+$\mathrm{nn_p}$)" a multi-level LM log-linearly combining word and phrase level LMs, after a linear interpolation between 4-gram LMs and an NNLM at both word and phrase level.

The results in table 3 show that the improvements from paraphrastic LMs and neural network LMs are largely additive. For example, the word level paraphrastic 4-gram LM outperformed the baseline 4-gram LM "w4g" by 0.3% absolute. The same improvement was retained when both LMs were combined with the word level NNLM, "w4g+$\mathrm{nn_w}$", as are shown in the 2nd and the 5th line of table 3 respectively. These results confirm the complementarity between paraphrastic LMs and neural network LMs as discussed in section 1. Consistent with the results shown in table 2, further improvements were obtained using multi-level LMs. The best performance was obtained using the paraphrastic multi-level LM shown in the bottom line of table 3, which used a three-way interpolation between the baseline LM, paraphrastic LM and neural network LM at both word and phrase level before a log-linear combination was performed. Using this LM, total error rate reductions of 0.9% absolute (9% relative) and 0.5% absolute (5% relative) were obtained over the baseline 4-gram word level LM "w4g" and the NNLM "w4g+$\mathrm{nn_w}$" respectively, both being statistically significant. The genre specific CER reductions over the baseline 4-gram LM "w4g" are 0.5% absolute (9% relative) for BN and 1.2% absolute (8% relative) for BC.

6. CONCLUSION AND RELATION TO PRIOR WORK

This paper investigated using statistical paraphrase approach to improve the context coverage and generalization of n-gram LMs for Mandarin Chinese broadcast speech recognition. The resulting paraphrastic LMs were then combined with word and phrase level neural network LMs. Significant error rate reductions of 5.0%-9.0% relative were obtained on a state-of-the-art LVCSR system trained on 1960 hours of speech and 1.5 billion words of text data. Together with earlier results published on a US English conversational telephone speech transcription task [15] for improving n-gram modelling only, the research presented in this paper demonstrates the proposed technique's cross genre generalization, scalability and complementarity with other modelling techniques. In contrast, previous research only investigated using manually derived expert word level synonym features [8, 10, 7, 3] to improve probability smoothing or word clustering. The statistical paraphrase generation based approach was not considered in any of these earlier works. Future research will focus on improving paraphrase extraction, modelling and directed paraphrasing for task and style adaptation.

7. REFERENCES

[1] I. Androutsopoulos & P. Malakasiotis (2010). A Survey of Paraphrasing and Textual Entailment Methods, *Journal of Artificial Intelligence Research*, 38:135-187, 2010.

[2] P. F. Brown et al. (1992). Class-based n-gram models of natural language. *Computational Linguistics* 18(4) pp.467-470.

[3] G. Cao, J-Y Nie & J. Bai (2005). Integrating word relationships into language models, in *Proc. ACM SIGIR2005*, pp. 298-305, Salvador, Brazil.

[4] Z. Dong & Q. Dong (2006). *HowNet And The Computation Of Meaning*, pp. 1-316, World Scientific, ISBN: 978-981-256-491-7.

[5] C. Fellbaum (1998) *WordNet: An Electronic Lexical Database*, MIT Press. Cambridge, MA.

[6] Z. Harris (1954). Distributional Structure, *Word*, 10(2):3 pp.146-162.

[7] R. Hoberman & R. Rosenfeld (2002). Using WordNet to Supplement Corpus Statistics [Online Document]. Available: http://www.cs.cmu.edu/∼roseh/Papers/wordnet.pdf, 2002.

[8] F. Jelinek, R. Mercer & S. Roukos (1990). Classifying words for improved statistical language models, in *Proc. IEEE ICASSP1990*, Vol. 1, pp. 621-624, Albuquerque, New Mexico.

[9] R. Kneser & H. Ney (1993), "Improved clustering techniques for class based statistical language modeling," in *Proc. EuroSpeech93'*, Berlin.

[10] R. Kneser & J. Peters (1997). Semantic clustering for adaptive language modeling, in *Proc. ICASSP1997*, Vol. 2, pp. 779-782, Munich.

[11] D. Lin & P. Pantel (2001). DIRT - Discovery of Inference Rules from Text, in *Proc. ACM SIGKDD2001*, pp.323-328, San Francisco, CA.

[12] X. Liu et al. (2010). Language Model Combination and Adaptation Using Weighted Finite State Transducers, in *Proc. IEEE ICASSP2010*, Dallas.

[13] X. Liu, M. J. F. Gales & P. C. Woodland (2012). Language Model Cross Adaptation for LVCSR System Combination, *Computer Speech and Language*, in press.

[14] X. Liu, J. L. Hieronymus, M. J. F. Gales & P. C. Woodland (2012). Syllable Language Models for Mandarin Speech Recognition: Exploiting Character Sequence Models, *Journal of the Acoustical Society of America*, in press.

[15] X. Liu, M. J. F. Gales & P. C. Woodland (2012). Paraphrastic Language Models, in *Proc. ISCA Interspeech2012*, Portland, Oregon.

[16] N. Madnani & B. Dorr (2010). Generating Phrasal and Sentential Paraphrases: A Survey of Data-Driven Methods, *Computational Linguistics*, Vol. 36, No. 3, 2010.

[17] M. Mohri. Finite-state transducers in language and speech processing. *Computational Linguistics*, 23:2, 1997.

[18] T. R. Niesler , E. W. D. Whittaker & P. C. Woodland (1998) Comparison Of Part-Of-Speech And Automatically Derived Category-Based Language Models For Speech Recognition, in *Proc. IEEE ICASSP1998*, Vol.1, pp. 177-180, Seattle, WA.

[19] J. Park, X. Liu, M. J. F. Gales & P. C. Woodland (2010). Improved Neural Network Based Language Modelling and Adaptation, in *Proc. ISCA Interspeech'10*, Makuhari.

[20] I. Oparin, M. Sundermeyer, H. Ney & J-L Gauvain (2012). Performance Analysis of Neural Networks in Combination with n-gram Language Models, in *Proc. IEEE ICASSP2012*, pp. 5005-5008, Kyoto.

[21] M. Pasca & P. Dienes (2005). Aligning needles in a haystack: Paraphrase acquisition across the Web, In *Proc. IJCNLP2005*, pp. 119-130, Jeju Island.

[22] R. Zens, F. Och & H. Ney (2002). Phrase-based Statistical Machine Translation, in *KI 2002: Advances in Artificial Intelligence*, pp. 35-56, 2002.

[23] H. Schwenk, "Continuous space language models", *Computer Speech and Language*, Vol. 21, 2007, pp. 492–518.