Paraphrastic 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, this paper presents a novel form of language model, the paraphrastic LM. A phrase level transduction model that is statistically learned from standard text data is used to generate paraphrase variants. LM probabilities are then estimated by maximizing their marginal probability. Significant error rate reductions of 0.5%-0.6% absolute were obtained on a state-of-the-art conversational telephone speech recognition task using a paraphrastic multi-level LM modelling both word and phrase sequences.

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. These paraphrase variants functionally represent different styles, dialects or other speaker specific characteristics. 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 [10, 12, 9, 5] 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 [7]. 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, this paper presents a novel form of language model, the paraphrastic LM. It allows a more flexible and general form of paraphrase modelling to be used at either the word, phrase or sentence level. A phrase level transduction model statistically learned from standard text data is used to generates multiple paraphrase variants. LM probabilities are then estimated by maximizing the marginal probability of these variants. The rest of the paper is organized as follows. Paraphrastic language models are introduced in section 2. A statistical n-gram phrase pair based paraphrase extraction scheme is proposed in section 3. Weighted finite state transducer (WFST) based paraphrase lattice generation is presented in section 4. In section 5 a range of paraphrastic LMs are evaluated on a state-of-the-art conversational telephone speech transcription task. Section 6 is the conclusion and possible future work.

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) in this paper. For a L word long word sequence $\mathcal{W} =< w_1, w_2, ..., w_i, ..., w_L >$ in the training data, rather than maximizing the surface word sequence log-probability $\ln P(\mathcal{W})$ as for conventional LMs, the marginal probability over all paraphrase variant sequences is maximized,

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

where

- $P(\boldsymbol{\psi}'|\mathcal{W}')$ 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;
- $P_{\text{PLM}}(\mathcal{W}')$ is paraphrastic LM probability to be estimated.

It can be shown that the sufficient statistics for a maximum likelihood (ML) estimation of $P_{\text{PLM}}(\mathcal{W}')$ are accumulated along

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

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

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

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 poorer 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 [17] 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

$$P(\tilde{w}|\tilde{h}) \;=\; \lambda P_{\mathsf{NG}}(\tilde{w}|\tilde{h}) + (1-\lambda) P_{\mathsf{PLM}}(\tilde{w}|\tilde{h}) \qquad (3)$$

where λ is the interpolation weight assigned to the conventional LM distribution $P_{\mathsf{NG}}(\cdot)$, and can be optimized on the perplexity of 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 criterion given in equation (1), where the word to phrase segmentation model $P(\boldsymbol{\psi}'|\mathcal{W}')$ is dropped,

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

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 [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 as 0.6 and 0.4 for word and phrase level LMs, and kept fixed for all experiments of this paper.

3. Paraphrase Phrase Pair Extraction

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 level, statistical paraphrase extraction schemes are needed.

Depending on the nature of the data being used, these techniques can be categorized into two major types [1, 15]. The first category uses comparable or parallel text data. Coarse grained alignment [2], or statistical machine translation based extraction methods are used to learn the paraphrastic relationship among words and phrases. As these methods assume a partial or complete semantic overlap between sentences, highly specialized training material is required. Hence, it is expensive to obtain and use on a large scale. The second category of techniques perform paraphrase pair extraction using standard text data [13, 19]. These are motivated by the *distributional similarity* theory [8], 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 following n-gram paraphrase induction algorithm is used to estimate the paraphrase model. The minimum and maximum phrase length are set as $L_{\min} = 1$ and $L_{\max} = 4$, and the left and right context length set as $L_N = 3$ and kept fixed for all experiments in this paper.

```
initialize phrase pair list $V = \{\}$;
initialize $n$-gram subsequence list $U = \{\}$;
**for** every sentence in training data **do**
    find and add all subsequences $<c_l, v, c_r>$ such that
    $L_{c_l} = L_N$, $L_{c_r} = L_N$ and $L_{\min} \le L_v \le L_{\max}$ into $U$.
**end for**
**for** every $<c_l, v, c_r>$ in $U$ **do**
    **for** every other $<c'_l, v', c'_r>$ in $U$ **do**
        **if** $c_l = c'_l$, $c_r = c'_r$ and $v \neq v'$ **then**
            **if** $<v \rightarrow v'>$ and $<v' \rightarrow v>$ not in $V$ **then**
                add phrase pairs $<v \rightarrow v'>$, $<v' \rightarrow v>$ to $V$;
            **end if**
            increase co-occurrence counts $C(v \rightarrow v')$
            and $C(v' \rightarrow v)$ both by 1;
        **end if**
    **end for**
**end for**
**for** every phrase pair $<v \rightarrow v'>$ in $V$ **do**
    estimate paraphrase prob $p(v'|v) = \frac{C(v \rightarrow v')}{\sum_{\bar{v}} C(v \rightarrow \bar{v})}$
**end for**
```

The above algorithm can be extended to incorporate additional useful information. For example, it is possible to build domain or style dependent paraphrastic LMs via a directed paraphrasing by restraining the choice of target phrases being used. In order to improve the grammaticality of paraphrase variants, syntactic constraints may be added. In common with other paraphrase induction methods, the above scheme can also produce phrase pairs that are non-paraphrastic, for example, producing antonyms. However, this is of less concern for language modelling, for which improving context coverage is the prime aim.

4. Paraphrase Lattice Generation

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) [16], WFST based paraphrase variant generation was used in this work, rather than designing special purpose decoding tools. For each training data sentence, the paraphrase word lattice $\mathcal{T}_{\mathcal{W}'}$ is generated using a sequence of WFST composition operations, before being projected onto the word sequence level and compressed via the determinization operation. This is given by

$$\mathcal{T}_{\mathcal{W}'} \;=\; \det\left(\pi_{\mathcal{W}'}\left(\mathcal{T}_{\mathcal{W}:\mathcal{W}} \circ \mathcal{T}_{\mathcal{W}:\boldsymbol{\psi}} \circ \mathcal{T}_{\boldsymbol{\psi}:\boldsymbol{\psi}'} \circ \mathcal{T}_{\boldsymbol{\psi}':\mathcal{W}'}\right)\right) \qquad (5)$$

where $\mathcal{T}_{\mathcal{W}:\mathcal{W}}$ is the transducer containing the original word sequence, $\mathcal{T}_{\mathcal{W}:\boldsymbol{\psi}}$ is the word to phrase segmentation transducer, $\mathcal{T}_{\boldsymbol{\psi}:\boldsymbol{\psi}'}$ the phrase to phrase paraphrase transducer and $\mathcal{T}_{\boldsymbol{\psi}':\mathcal{W}'}$ the phrase to word transducer. $\circ$, $\det(\cdot)$ and $\pi(\cdot)$ denote the WFST composition, determinization and projection operations.

An example of a word to phrase segmentation transducer is shown in figure 1 (a), which can generate three phrases, a single word phrase $v_1 : w_1$, a two word phrase $v_2 : w_2 w_3$ and a three word one $v_3 : w_4 w_5 w_6$. Here <e> denotes the null symbol. As discussed in section 2, both the phrase to word, and word to phrase segmentation models are considered non-informative in this work. The phrase to word transducer can thus be obtained by taking the word to phrase transducer's inverse (swapping input and output symbols). An example of a phrase to phrase paraphrase model is shown in figure 1 (b), where phrase v_1 is transformed into either v_4 with a probability of 0.7 or v_5 at 0.3 (0.36 and 1.22 as negated log prob), while phrase v_2 is paraphrased into v_6 or v_7 with the associated costs of 1.6 or 0.219.

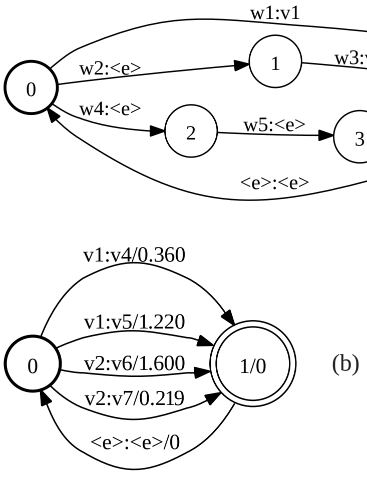

Figure 1: Example WFST representation of (a) word to phrase segmentation model and (b) phrase to phrase paraphrase model.

Using the above WFST based decoding approach and a paraphrase model trained on 545 million words of conversational data, for an example sentence "*And I generally prefer*", the following paraphrase variants are among those generated: "*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* " and "*Um I wish*". As the paraphrase extraction method presented in section 3 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.

5. Experiments and Results

In this section performance of various paraphrastic language models are evaluated on the CU-HTK LVCSR system for conversational telephone speech (CTS) used in the 2004 DARPA EARS evaluation. The acoustic models were trained on approximately 2000 hours of Fisher conversational speech released by the LDC. A 59k 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, conversational side level normalized PLP features, and an interpolated 3-gram word level baseline LM were used. A detailed description of the baseline system can be found in [6]. The 3 hour **dev04** data, which includes 72 Fisher conversations, was used as a test set. 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 1.0 billion words from 8 difference text sources. The two text sources with the highest interpolation weights, the LDC Fisher acoustic transcriptions, **Fisher**, of 20 million words (0.6), and the University Washington conversational web data [4], **UWWeb** of 525 million words (0.2), were used to build various language models. These LMs are then used for lattice rescoring and word error rate (WER) performance evaluation. Information on corpus size, paraphrase extraction schemes used and the number of phrase pairs extracted from the these two text sources, as well as WordNet, are given in table 1. Using the automatic n-gram paraphrase extraction scheme presented in section 3, a total of 90k and 2.9M phrase pairs were extracted from the **Fisher** and **UWWeb** data respectively. The expert semantic labelling by WordNet, including synonyms, hypernyms, hyponyms and pertainyms, were used to generate 480k paraphrase phrase pairs.

Source	**Size**	**Extraction**	**# Phrase Pairs**
WordNet	-	Expert	480k
Fisher	20M	Automatic	90k
UWWeb	525M	Automatic	2.9M

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

WER performance of various LMs trained using the **Fisher** data only are shown in table 2 for **dev04**. The first three baseline LMs are non-paraphrastic. The word level 4-gram baseline LM "w4g" gave a WER of 17.6%. When further interpolated with a class based LM of 1000 automatically derived word clusters[11], the "w4g+clslm" model reduces the error rate by 0.2% absolute. The third baseline LM in table 2 is a multi-level LM, "w4g $\circ$ p4g", which incorporates phrase level linguistic constraints by log-linearly combining the word and phrase level 4-gram LMs. It was constructed by adding a total of 16k distinct multi-word phrases found in the **Fisher** data generated paraphrase phrase table to the baseline 59k word list, and trained on the phrase level text data obtained using a longest available word to phrase segmentation. This is similar to the method used in [18]. As discussed in section 2, word level lattices need to be first converted to phrase level lattices when using the multi-level LM. This was implemented using a WFST composition between the word level lattice with the phrase level segmentation transducer shown in figure 1(a). After the log-linear combination between word and phrase level LMs is performed, the resulting phrase level lattices are converted back to word level again via a WFST composition with the phrase to word transducer, to obtain the 1-best word level hypothesis for WER evaluation. By adding additional phrase level features, this multi-level LM gives a small improvement of 0.1% absolute over the word level 4-gram baseline LM.

In contrast, the comparable word level paraphrastic 4-gram LM, shown in the 4th line of table 2, using the paraphrase phrase

LM	Paraphrastic	dev04
w4g	×	17.6
w4g+clslm		17.4
w4g $\circ$ p4g		17.5
w4g	√	17.2
w4g $\circ$ p4g		17.0

Table 2: Performance of LMs trained using the **Fisher** data only for **dev04**. “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 $\circ$ p4g” a multi-level LM log-linearly combining word and phrase level 4-gram LMs.

pairs extracted from the **Fisher** training data itself and WordNet, as given in table 1, outperformed the word level baseline 4-gram LM, and the class LM baseline, by 0.4% and 0.2% absolute respectively. Similarly when using the paraphrastic multi-level LM, shown in the last line of table 2, a significant WER reduction of 0.5% was obtained over the baseline non-paraphrastic multi-level LM shown in the 3rd line of table 2. The overall improvement over the word level 4-gram baseline LM is 0.6% absolute, which is also statistically significant. It was also found that adding the paraphrases extracted from WordNet gave only a marginal improvement over using only those automatically learned from the **Fisher** data. For example, a comparable paraphrastic multi-level LM derived using only the 90k phrase pairs obtained from **Fisher** data gave a very similar WER performance of 17.1%. This is expected as the **Fisher** corpus provides the in-domain data for the CTS task, while the expert paraphrases of WordNet are more task independent.

LM	Paraphrastic	dev04
w4g	×	16.7
w4g+clslm		16.5
w4g $\circ$ p4g		16.5
w4g	√	16.4
w4g $\circ$ p4g		16.2

Table 3: Performance of LMs trained using **Fisher** and **UWWeb** data on **dev04**. Naming convention same as table 2.

The same trend can also be found in a set of experiments conducted on a larger LM training set up where the 525M word **UWWeb** data is also used in LM training as a second source via a linear interpolation with the **Fisher** data trained LM. As expected, adding this data source significantly improved the performance of three non-paraphrastic baseline LMs by 0.9%-1.0% absolute, compared with the results shown in the first three lines of table 2. The word level paraphrastic 4-gram LM, as is shown in the 4th line of table 3, using the paraphrase phrase pairs extracted from all three data sources given in table 1, outperformed the word level baseline LM by 0.3% absolute. When using the paraphrastic multi-level LM, as is shown in the last line of table 3, an overall significant WER reduction of 0.5% absolute was obtained over the word level 4-gram baseline LM. It also outperformed a word level 4-gram baseline LM trained using twice the amount of data, 1.0 billion words, with 6 more text sources in addition to **Fisher** and **UWWeb**, by 0.2%.

6. Conclusion

Paraphrastic language models were investigated in this paper. Significant error rate reductions of 0.5%-0.6% absolute were obtained on a state-of-the-art large vocabulary speech recognition task. Experimental results suggest the proposed method is effective in improving LM context coverage and generalization performance, and thus may be useful for speech recognition. Future research will focus on using more data in paraphrastic LM training, improving paraphrase pair extraction, modelling method 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] R. Barzilay & L. Lee (2003). Learning to paraphrase: An unsupervised approach using multiple-sequence alignment, in *Proc. of HLT-NAACL 2003*, pp 16-23, Edmonton.

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

[4] I. Bulyko, M. Ostendorf & A. Stolcke. Getting More Mileage from Web Text Sources for Conversational Speech Language Modeling using Class-Dependent Mixtures, in *Proc. HLT’03*, Edmonton.

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

[6] G. Evermann et al. (2004). Training LVCSR Systems on Thousands of Hours of Data, in *Proc. ICASSP2005*, Philadelphia.

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

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

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

[10] 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.

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

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

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

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

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

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

[17] 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.

[18] M. Padmanabhan et al. (1998). Speech Recognition Performance on a Voicemail Transcription Task, In *Proc. IEEE ICASSP1998*, Vol. 2 pp. 913-916, Seattle, WA.

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