Paraphrastic Language Models

Xunying Liu
CUED NST Team

April 24th 2012

Cambridge University Engineering Department

Introduction

- **Multiple sentences represent the same meaning in natural languages.**

- **Meaning to surface word sequence mapping often one-to-many:**
  - carrying identical or similar semantic information, paraphrastic to each other;
  - but using different syntactic, lexical and morphological rules in generation;
  - representing e.g., different styles, dialects or other speaker characteristics;
  - poor coverage if only modelling surface form using, e.g., n-gram LMs.

- **Alternatively possible to directly model paraphrase variants:**
  - should improve LM context coverage and generalization;
  - existing methods focus on modelling word level synonyms only;
  - often use expert semantic labelling, expensive to manually derive;
  - **more complicated paraphrase mapping at higher level is preferred;**
  - **automatic paraphrase extraction techniques are required.**

Paraphrastic Language Models

- **Flexibly model word/phrase/sentence level paraphrase mapping:**
  - phrase level transduction model generates multiple paraphrase variants;
  - language model probabilities estimated in paraphrased domain;
  - by maximizing marginal probability of paraphrase sequences.
- **Modelling alternative expressions of same meaning:**
  - improves domain, context coverage and generalization;
  - possible to build domain or style independent language models;
  - directed paraphrasing can also produce domain dependent LMs.
- **Appropriate paraphrase pair extraction scheme is important:**
  - impractical to obtain expert semantic labelling on phrase level;
  - **automatic paraphrase pair extraction scheme is preferred.**

Paraphrase Pair Extraction

- **Paraphrase pair extraction using standard text data:**
  - **distributional similarity** based paraphrase pair induction scheme;
  - phrase pairs often sharing same L/R contexts considered paraphrases;
  - standard text data in large amounts can be used to improve coverage.

- **Paraphrase pair extraction using comparable or parallel text data:**
  - using coarse grained alignment, or SMT based extraction methods;
  - assuming partial or complete semantic overlap between sentences;
  - highly specialized training material, expensive to obtain in large quantity.

- **Current research focused on using standard text data:**
  - n**-gram phrase based approach** extracts variable length n-gram pairs;
  - possible to produce non-paraphrastic phrase pairs, e.g., antonyms;
  - but for language modelling improving context coverage is the prime aim.

Paraphrase Pair Extraction for Conversational Data

- **LDC Fisher data (20M) and UW conversational web data (525M) used:**
  - min length 1 word, max length 4 words, 3 word left and right context;
  - expert paraphrases extracted from WordNet also used to improve coverage.

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

- **WFST based paraphrase lattice generation for LM training:**
  - efficient representation of phrase level paraphrase model;
  - well defined search algorithms available, no special purpose decoder required.

Paraphrase Pair Extraction for Conversational Data (Example)

Orignial 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 ... ...

Performance of Paraphrastic Language Models

- **Adapted 2000 hour Fisher data trained PLP MPE acoustic models:**
  - 20M words Fisher data transcription trained LMs for lattice rescoring;
  - 4-gram word and phrase level LMs intersected to construct multi-level LMs;
  - word, class-based, multi-level baseline and paraphrastic LMs evaluated.

LM	Paraphrastic	dev04
w4g	$\times$	17.6
w4g+clslm		17.4
w4g $\circ$ p4g		17.5
w4g	$\surd$	17.2
w4g $\circ$ p4g		17.0

- **Consistent performance gains over word and multi-level baseline LMs.**
- **WER reduction of 0.6% abs. (3.4% rel.) over word 4-gram baseline.**

Performance of Paraphrastic Language Models (Cont)

- **LMs trained using both LDC Fisher and UW web data evaluated:**
  - 525M words of UW conversational web data used as additional text source;
  - word, class-based, multi-level baseline and paraphrastic LMs evaluated.

LM	Paraphrastic	dev04
w4g	×	16.7
w4g+clslm	×	16.5
w4g ∘ p4g	×	16.5
w4g	√	16.4
w4g ∘ p4g	√	16.2

- **Performance gains consistent with experiments using Fisher data only.**
- **WER reduction of 0.5% abs. (3% rel.) over word 4-gram baseline.**

Conclusion and Future Work

- **Paraphrastic language models useful for speech recognition.**
- **Improving LM generalization and context coverage.**
- **Future research will focus on:**
  - using more data in paraphrastic LM training;
  - improving paraphrase pair extraction;
  - using more complicated forms of paraphrase domain LMs;
  - directed paraphrasing in LM training for task and style adaptation.