Cross-domain Paraphrasing For Improving Language Modelling Using Out-of-domain Data

Xunying Liu

CUED NST Team

May 23th 2013

Cambridge University Engineering Department

Outline

- **This talk summarizes recent research on using cross-domain paraphrasing to improve language modelling using out-of-domain data.**
- **Main objectives to achieve under NST:**
  - improving domain and context coverage for language models;
  - structured modelling of domain commonalities and variabilities;
  - minimising demand for in-domain training data;
  - to be rapidly deployable for a new target domain or situation.
- **Initial results on two state-of-the-art LVCSR tasks:**
  - conversational telephone speech transcription;
  - multi-genre media archive transcription.

Introduction

- **Language modelling using data from different domains is difficult:**
  - more challenging when only limited amounts of in-domain data is available;
  - out-of-domain data are often available in large quantities;
  - but they are less useful due to domain mismatch.
- **Variability in generation rules significantly alters surface word sequence:**
  - thus introduces mismatch against other related data generated via alternative realizations associated with characteristics of e.g. a different domain.
- **Directly modelling out-of-domain data or std. data/model combination:**
  - ignores domain commonalities/specialties, inefficient use of OOD data;
  - results in poor in-domain data coverage using, e.g., n-gram LMs.
- **Alternatively possible to structurally exploit domain independent and dependent characteristics of in-domain and out-of-domain training data.**

Cross-domain Paraphrasing

- **Out-of-domain data contains rich domain independent generation rules:**
  - representing, e.g. paraphrastic relationships between words/phrases/sentences;
  - cross-domain paraphrasing expands limited in-domain training material;
  - should produce richer context coverage of in-domain data.
- **In-domain data viewed as paraphrases of related out-of-domain data:**
  - assuming some overlap in topic and meaning with in-domain data;
  - generated using alternative domain specific realization rules representing;
  - e.g. disfluency/non-grammaticality/informal style of conversational data;
  - cross-domain paraphrasing gives “in-domain like” paraphrases of OOD data;
  - more effective use of out-of-domain data to improve in-domain LM coverage.
- **Cross-domain paraphrasing leverages strengths of ID and OOD data:**
  - wider coverage than data/model combi. or in-domain paraphrases only;
  - **xdomain paraphrases used to improve paraphrastic LM performance.**

Paraphrastic Language Models

- **Flexibly model word/phrase/sentence level paraphrase mapping:**
  - phrase level paraphrase 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:**
  - intuitive and interpretable smoothing mechanism;
  - improves domain, context coverage and generalization;
  - can incorporate additional phrase level linguistic constraints.
- **Appropriate paraphrase pair extraction scheme is important:**
  - impractical to obtain expert semantic labelling on phrase level;
  - **automatic paraphrase pair extraction scheme is preferred.**

Paraphrastic Language Models (cont)

- **n-gram phrase based paraphrase learning from standard text:**
  - std. text in large amounts with no semantic labelling to improve coverage;
  - phrases often sharing same L/R contexts more likely to be paraphrases;
  - syntactic constraints may be added to improve grammaticality.
- **WFST based efficient training data paraphrase lattice generation:**
  - well defined algorithms available, no special purpose decoder;
  - lattice fwdbwd pass generates statistics for paraphrastic LM training.
- **Linear interpolation with standard n-gram LMs to achieve:**
  - good balance between context coverage and discrimination.

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

- **Paraphrase learning/generation within same domain in previous work:**
  - **possible to further improve performance via xdomain paraphrasing.**

Cross-domain Paraphrastic Language Models

- **Both in and cross domain paraphrases generated for PLM training.**
- **Generating more ID data using OOD data trained paraphrase model:**
  - retaining sentential structure, topic coverage and semantics of ID data;
  - exploiting additional rich domain independent paraphrases in OOD data.
- **Reducing OOD data domain mismatch using ID paraphrase model:**
  - transforms OOD data into "in-domain like" data via a directed paraphrasing;
  - by restraining target paraphrases found only in the in-domain data;
  - injecting domain specific characteristics, e.g. disfluency and informal style.
- **Linear interpolation with standard n-gram LMs:**
  - In-domain and cross-domain paraphrases used to build separate PLMs.

$$P(\tilde{w}|\tilde{h}) = \lambda_{\text{NG}} P_{\text{NG}}(\tilde{w}|\tilde{h}) + \lambda_{\text{PLM}}^{\text{ID}} P_{\text{PLM}}^{\text{ID}}(\tilde{w}|\tilde{h}) + \lambda_{\text{PLM}}^{\text{XD}} P_{\text{PLM}}^{\text{XD}}(\tilde{w}|\tilde{h})$$

In/Cross-domain Paraphrasing for Conversational Data (e.g.)

Original sentence: And I generally prefer

In-domain Paraphrases: 545M words ID data, 3M paraphrase pairs.

And I really like	I mean I would like
I guess I generally like	I think I need
So I appreciate	You know I just want
'Cause I love	Well I prefer
Um I wish	Yeah I just prefer

... ...

Cross-domain Paraphrases: 490M words BN data, 2.9M paraphrase pairs.

Actually I love	But largely I you know prefer
And I honestly generally hope	And possibly I choose
Or I just really intend	And quite frankly I probably prefer
Seemed like I always very much like	And I tend to probably prefer
And more remarkably I usually prefer	And personally I generally want

... ...

In/Cross-domain Paraphrasing for Broadcast News (e.g.)

Original sentence: Economy is a big problem for the Bush administration

In-domain Paraphrases: 490M words ID data, 2.9M paraphrase pairs.

Economy is an uphill battle very much for the White House
Economy will be a main problem for the United States
Economy is a real challenge for the administration
Economy represents a big trouble for the Bush presidency
Economy constitutes a large problem for the Bush government
... ...

Cross-domain Paraphrases: 545M word conv. data, 3M paraphrase pairs.

Economy is a heck of a uh problem for the president
Economy is a big big deal for the president
Economy is an awful big problem for I mean president
Economy is like a big problem for uh the Bush administration
Economy of course that's a big problem I think the president
... ...

Experiments on Conversational Telephone Speech

- **Adapted 2000 hour Fisher data trained PLP MPE acoustic models:**
  - 545M words ID (LDC Fisher + UWWeb) data, 490M words OOD BN text;
  - 5.9M statistically derived and 480k expert (WordNet) paraphrase pairs;

LM	Paraphrastic	Cross-domain	Miss Rate(%)	
			3g	4g
w4g	×	×	17.9	49.4
	√	×	14.3	42.9
	√	√	13.1	39.8

- **Consistent n-gram coverage improvements over baseline paraphrastic LM trained without using cross-domain generated paraphrases.**
- **19%-27% 3/4-gram miss rate reduction (30% reduction from xdomain paraphrasing) over baseline 4-gram LM built using model interpolation.**

Experiments on Conversational Telephone Speech (cont)

- **4-gram word/phrase level LMs intersected to construct multi-level LMs.**
- **word, class-based, multi-level baseline and paraphrastic LMs evaluated.**

LM	Paraphrastic	Cross-domain	dev04
w4g w4g+clslm w4g $\circ$ p4g	$\times$	$\times$	16.6 16.4 16.4
w4g w4g $\circ$ p4g	$\surd$	$\times$	16.3 16.1
w4g w4g $\circ$ p4g	$\surd$	$\surd$	16.2 16.0

- **Consistent improvements over word and multi-level baseline PLMs.**
- **WER reduction of 0.6% abs. (4% rel.) over word 4-gram baseline.**

Experiments on Media Archive Data

- **21 hour BBC media archive data, adapted PLP MPE acoustic models:**
  - 250k words ID BBC transcripts, 490M words OOD BN data;
  - 2.9M statistically derived and 480k expert (WordNet) paraphrase pairs.

LM	Paraphrastic	Cross-domain	bbcdev
w4g w4g+clslm w4g $\circ$ p4g	$\times$	$\times$	31.3 31.2 31.0
w4g w4g $\circ$ p4g	$\surd$	$\times$	30.9 30.7
w4g w4g $\circ$ p4g	$\surd$	$\surd$	30.7 30.5

- **15%-21% 3/4-gram miss rate reduction (35% reduction from xdomain paraphrasing) over baseline 4-gram LM built using model interpolation.**
- **WER reduction of 0.8% abs. (3% rel.) over word 4-gram baseline.**

Conclusion

- **Xdomain paraphrases improve LM domain coverage and generalization:**
  - naturally link language generation with language modelling;
  - structured modelling of domain independent/dependent characteristics;
  - generate in-domain training data from rich out-of-domain data;
  - minimise demand for in-domain training data;
  - reduce domain mismatch and more efficient use of out-of-domain data;
  - can be rapidly deployed in under-resourced new domains.