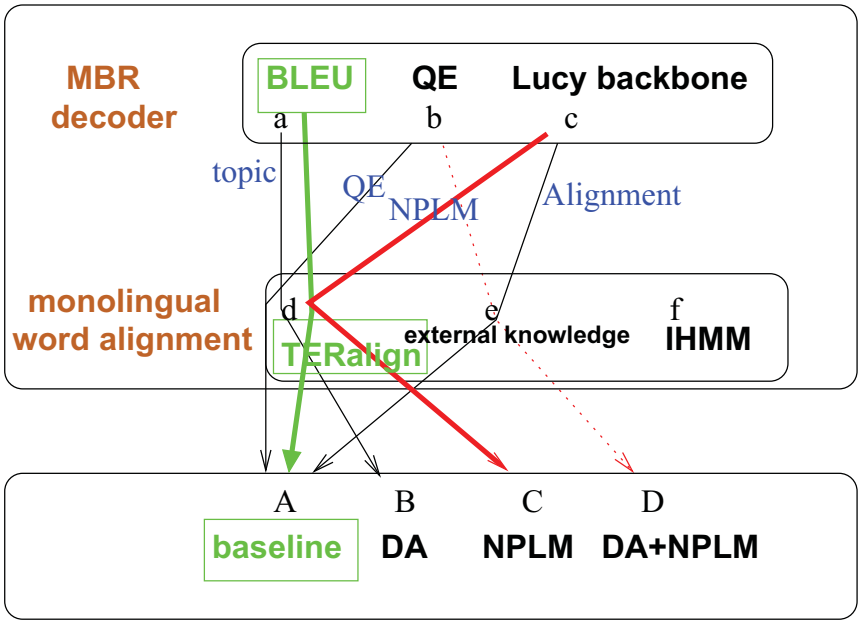
Neural Probabilistic Language Model for System Combination

Tsuyoshi Okita
Dublin City University

Centre for Next Generation Localisation
NDP National Development Plan 2007 - 2013
sfi science foundation ireland
DCU Dublin City University
UCD Dublin University College Dublin
University of Limerick
Trinity College Dublin

DCU-NPLM Overview



System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network
     - Select a backbone by Minimum-Bayes Risk (MBR) decoder (with MERT tuning)

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

Centre for Next Generation Localisation

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network
     - Select a backbone by Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
     - Run monolingual word aligner

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

Centre for Next Generation Localisation

System Combination Overview

- System combination [Matusov et al., 05; Rosti et al., 07]
- Given: Set of MT outputs
  1. Build a confusion network
     - Select a backbone by Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
     - Run monolingual word aligner
  2. Run monotonic (consensus) decoder (with MERT tuning)

- We focus on three technical topics
  1. Minimum-Bayes Risk (MBR) decoder (with MERT tuning)
  2. Monolingual word aligner
  3. Monotonic (consensus) decoder (with MERT tuning)

System Combination Overview

Input 1	they are normally on a week .
Input 2	these are normally made in a week .
Input 3	este himself go normally in a week .
Input 4	these do usually in a week .

⇓ **1. MBR decoding**

Backbone(2)	these are normally made in a week .

⇓ **2. monolingual word alignment**

Backbone(2)	these	are	normally	made	in	a	week .
hyp(1)	they$_S$	are	normally	*****$_D$	on$_S$	a	week .
hyp(3)	este$_S$	himself$_S$	go$_S$	normally$_S$	in	a	week .
hyp(4)	these	*****$_D$	do$_S$	usually$_S$	in	a	week .

⇓ **3. monotonic consensus decoding**

Output	these	are	normally	*****	in	a	week .

1. MBR Decoding

1. Given MT outputs, choose 1 sentence.

$$\begin{aligned}
\hat{E}_{best}^{MBR} &= \text{argmin}_{E' \in \mathcal{E}} R(E') \\
&= \text{argmin}_{E' \in \mathcal{E}} \sum_{E' \in \mathcal{E}_E} L(E, E') P(E|F) \\
&= \text{argmin}_{E' \in \mathcal{E}} \sum_{E' \in \mathcal{E}_E} (1 - BLEU_E(E')) P(E|F) \\
&= \text{argmin}_{E' \in \mathcal{E}} \\
&\quad \left[\mathbf{1} - \left[\begin{array}{cccc} B_{E_1}(E_1) & B_{E_2}(E_1) & B_{E_3}(E_1) & B_{E_4}(E_1) \\ B_{E_1}(E_2) & B_{E_2}(E_2) & B_{E_3}(E_2) & B_{E_4}(E_2) \\ & \cdots & \cdots & \\ B_{E_1}(E_4) & B_{E_2}(E_4) & B_{E_3}(E_4) & B_{E_4}(E_4) \end{array} \right] \right] \left[\begin{array}{c} P(E_1|F) \\ P(E_2|F) \\ P(E_3|F) \\ P(E_4|F) \end{array} \right]
\end{aligned}$$

1. MBR Decoding

Input 1	they are normally on a week .
Input 2	these are normally made in a week .
Input 3	este himself go normally in a week .
Input 4	these do usually in a week .

$$
\begin{aligned}
&= \text{argmin} \left[\mathbf{1} - \begin{bmatrix} 1.0 & 0.259 & 0.221 & 0.245 \\ 0.267 & 1.0 & 0.366 & 0.377 \\ & \dots & \dots & \\ 0.245 & 0.366 & 0.346 & 1.0 \end{bmatrix} \right] \begin{bmatrix} 0.25 \\ 0.25 \\ 0.25 \\ 0.25 \end{bmatrix} \\
&= \text{argmin}\,[0.565, 0.502, 0.517, 0.506] \\
&= (\mathit{Input}2)
\end{aligned}
$$

Backbone(2)	these are normally made in a week .

2. Monolingual Word Alignment

- TER-based monolingual word alignment
  - Same words in different sentence are aligned
  - Proceeded in a pairwise manner: Input 1 and backbone, Input 3 and backbone, Input 4 and backbone.

Backbone(2)	these	are	normally	made	in	a	week .
hyp(1)	$they_S$	are	normally	$*****_D$	on_S	a	week .
Backbone(2)	these	are	normally	made	in	a	week .
hyp(3)	$este_S$	$himself_S$	go_S	$normally_S$	in	a	week .
Backbone(2)	these	are	normally	made	in	a	week .
hyp(4)	these	$*****_D$	do_S	$usually_S$	in	a	week .

3. Monotonic Consensus Decoding

- Monotonic consensus decoding is limited version of MAP decoding
  - monotonic (position dependent)
  - phrase selection depends on the position (local TMs + global LM)

$$
\begin{aligned}
e_{best} &= \arg\max_{e} \prod_{i=1}^{I} \phi(i|\bar{e}_i) p_{LM}(e) \\
&= \arg\max_{e} \{\phi(1|\text{these})\phi(2|\text{are})\phi(3|\text{normally})\phi(4|\emptyset)\phi(5|\text{in}) \\
&\quad \phi(6|\text{a})\phi(7|\text{week}) p_{LM}(e), \ldots\} \\
&= \text{these are normally in a week} \qquad (1)
\end{aligned}
$$

1 ||| these ||| 0.50	2 ||| are ||| 0.50	3 ||| normally ||| 0.50
1 ||| they ||| 0.25	2 ||| himself ||| 0.25	...
1 ||| este ||| 0.25	2 ||| $\emptyset$ ||| 0.25	...

Centre for Next Generation Localisation

Table Of Contents

Overview

1. N-gram language model
2. Smoothing methods for n-gram language model [Kneser and Ney, 95; Chen and Goodman, 98; Teh, 06]
   - Particular interest on unseen data
3. Neural probabilistic language model (NPLM)[Bengio, 00;Bengio et al., 2005]

- Perplexity: $1 < 2 < 3$

N-gram Language Model

- N-gram Language Model $p(W)$ (where W is a string $w_1, \ldots, w_n$)
  - $p(W)$ is the probability that if we pick a sentence of English words at random, it turns out to be W.
  - Markov assumption
    - Markov chain:
      $p(w_1, \ldots, w_n) = p(w_1)p(w_2|w_1) \ldots p(w_n|w_1, \ldots, w_{n-1})$
    - History under m words:
      $p(w_n|w_1, \ldots, w_{n-1}) \approx p(w_n|w_{n-m}, \ldots, w_{n-1})$
- Perplexity (This measure is used when one tries to model an unknown probability distribution p, based on a training sample that was drawn from p.)
  - Given a proposed model q, the perplexity, defined as

    $$2^{\sum_{i=1}^{N} \frac{1}{N} \log_2 q(x_i)},$$

    suggests how well it predicts a separate test sample $x_1, \ldots, x_N$ also drawn from p.

Language Model Smoothing(1)

- Motivation: unseen n-gram problem
  - An n-gram which was not appeared in the training set may appear in the test set.
    1. The probability of n-grams in training set is too big.
    2. The probability of unseen n-grams is zero.
  - (Some n-grams which will be reasonably appeared based on the lower- / higher-order n-grams may not appeared in the training set.)
- Smoothing method is
  1. to adjust the empirical counts that we observe in the training set to the expected counts of n-grams in previously unseen text.
  2. to estimate the expected counts of unseen n-grams included in test set. (Often no treatment)

Language Model Smoothing (2)

maximum likelihood	$P(w_i\|w_{i-1}) = \frac{c(w_{i-1}w_i)}{\sum_w c(w_{i-1}w)}$
add one	$P(w_i\|w_{i-1}) = \frac{c(w_{i-1}w_i)+1}{\sum_w c(w_{i-1}w)+v}$
absolute discounting	$P(w_i\|w_{i-1}) = \frac{c(w_{i-1}w_i)-D}{\sum_w c(w_{i-1}w)}$
Kneser-Ney	$P(w_i\|w_{i-1}) = \frac{c(w_{i-1}w_i)-D}{\sum_w c(w_{i-1}w)}$, $\alpha(w_i)\frac{N_{1+}(\bullet w)}{N_{1+}(w_{i-1}w)}$
interpolated modified KN	$P(w_i\|w_{i-1}) = \frac{c(w_{i-1}w_i)-D_i}{\sum_w c(w_{i-1}w)} + \beta(w_i)\frac{N_{1+}(\bullet w)}{N_{1+}(w_{i-1}w)}$ $D_1 = 1 - 2YN_2/N_1, D_2 = 2 - 3YN_3/N_2$ $D_{3+} = 3 - 4YN_4/N_3, Y = N_1/(N_1 + 2N_2)$
hierarchical PY	$P(w_i\|w_{i-1}) = \frac{c(w_{i-1}w_i)-d\cdot t_{hw}}{\sum_w c(w_{i-1}w)+\theta} + \delta(w_i)\frac{N_{1+}(\bullet w)}{N_{1+}(w_{i-1}w)}$ $\delta(w_i) = \frac{\theta+d\cdot t_{h\cdot}}{\theta+\sum_w c(w_{i-1}w)}$

Table: Smoothing Method for Language Model

Neural Probabilistic Language Model

- Learning representation of data in order to make the probability distribution of word sequences more compact
- Focus on similar semantical and syntactical roles of words.
  - For example, two sentences
    - *"The cat is walking in the bedroom"* and
    - *"A dog was running in a room"*
  - Similarity between (the, a), (bedroom, room), (is, was), and (running, walking).
- Bengio's implementation [00].
  - Implemention using multi-layer neural network.
  - 20% to 35% better perplexity than the language model with the modified Kneser-Ney methods.

Neural Probabilistic Language Model (2)

- to capture the semantically and syntactically similar words in a way that a latent word depends on the context (Below ideal situation)

a	japanese	electronics	executive	was	kidnapped
the	u.s.	tabacco	director	is	abducted
its	german	sales	manager	were	killed
one	british	consulting	economist	be	found
	russian		spokesman	are	abduction

System Combination with NPLM Plain

- The task of Word Sense Disambiguation using NPLM:

$$P(\text{synset}_i|\text{features}_i,\theta) \quad = \quad \frac{1}{Z(\text{features})}\prod_m g(\text{synset}_i,k)^{f(\text{feature}_i^k)}$$

  - k ranges over all possible features,
  - $f(\text{feature}_i^k)$ is an indicator function whose value is 1 if the feature exists, and 0 otherwise,
  - $g(\text{synset}_i,k)$ is a parameter for a given synset and feature,
  - θ is a collection of all these parameters in $g(\text{synset}_i,k)$,
  - Z is a normalization constant.
- We do reranking.

System Combination with NPLM Plain (2)

(a)	*the Government wants to limit the torture of the " witches " , as it published in a brochure*
(b)	the Government wants to limit the torture of the " witches " , as it published in the proceedings
(a)	*the women that he " return " witches are sent to an area isolated , so that they do not hamper the rest of the people .*
(b)	the women that he " return " witches are sent to an area eligible , so that they do not affect the rest of the country .

Table: Table includes two examples of plain paraphrase.

System Combination with NPLM Plain (3)

Given: For given testset g, prepare N translation outputs $\{s_1, \ldots, s_N\}$ from several systems, trained NPLM.

Step 1: Paraphrases the translation outputs $\{s_1, \ldots, s_N\}$ replaced with alternative expressions (or paraphrases).

Step 2: Augment the sentences of translation outputs prepared in Step 2.

Step 3: Run the system combination module.

System Combination with NPLM Dep (1)

- Noise is not negligible! (NPLM trained on small corpus)
- Removed by modified dependency score [Owczarzak et al., 07]
  - If we add paraphrases and the resulted sentence has a higher score in terms of the modified dependency score.
  - If the resulted score decreases, we will not add them (=noise).
- Naive approach (= MBR Decoding)
  - If we add paraphrases and the resulted sentence does not have a very bad score, we add these paraphrases since these paraphrase are not very bad (*naive* way).
  - Pairwise manner.

System Combination with NPLM Dep (2)

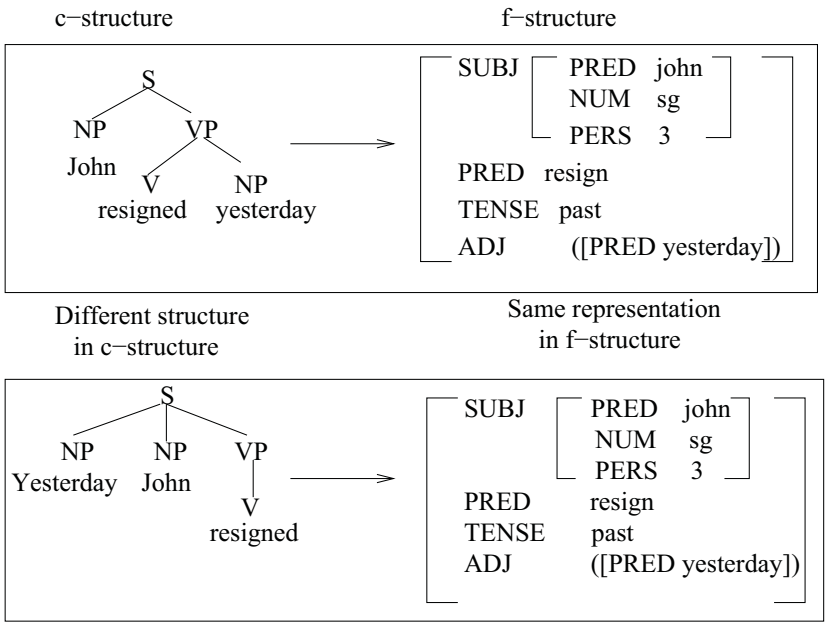


Figure: By the modified dependency score [Owczarzak,07], the score of these two sentences, "John resigned yesterday" and "Yesterday John resigned", are the same.

System Combination with NPLM Dep (3)

system	translation output	precision	recall	F-score
s1	these do usually in a week .	0.080	0.154	0.105
s2	these are normally made in a week .	0.200	0.263	0.227
s3	they are normally in one week .	0.080	0.154	0.105
s4	they are normally on a week .	0.120	0.231	0.158
ref	the funding is usually offered over a one-week period .			

Table: An example of modified dependency score

Experimental Settings

- ML4HMT-2012 datasets: four translation outputs (*s1* to *s4*) which are MT outputs by two RBMT systems, APERTIUM and LUCY, PB-SMT (MOSES) and HPB-SMT (MOSES).
- Tuning data 20,000 sentence pairs, and test data 3,003 sentence pairs.

Results

	NIST	BLEU	METEOR	WER	PER
s1	6.4996	0.2248	0.5458641	64.2452	49.9806
s2	6.9281	0.2500	0.5853446	62.9194	48.0065
s3	7.4022	0.2446	0.5544660	58.0752	44.0221
s4	7.2100	0.2531	0.5596933	59.3930	44.5230
NPLM plain	7.6041	0.2561	0.5593901	56.4620	41.8076
NPLM dep	7.6213	0.2581	0.5601121	56.1334	41.7820
BLEU-MBR	7.6846	0.2600	0.5643944	56.2368	41.5399
modDep precision	7.6670	0.2636	0.5659757	56.4393	41.4986
modDep recall	7.6695	0.2642	0.5664320	56.5059	41.5013
modDep Fscore	7.6695	0.2642	0.5664320	56.5059	41.5013

Results

	NIST	BLEU	METEOR	WER	PER
BLEU-MBR	7.6846	0.2600	0.5643944	56.2368	41.5399
min ave TER-MBR	7.6231	0.2638	0.5652795	56.3967	41.6092
DA	7.7146	0.2633	0.5647685	55.8612	41.7264
QE	7.6846	0.2620	0.5642806	56.0051	41.5226
s2 backbone	7.6371	0.2648	0.5606801	56.0077	42.0075
modDep precision	7.6670	0.2636	0.5659757	56.4393	41.4986
modDep recall	7.6695	0.2642	0.5664320	56.5059	41.5013
modDep Fscore	7.6695	0.2642	0.5664320	56.5059	41.5013

	modDep precision	modDep recall	modDep Fscore
average s1	0.244 (586)	0.208	0.225
average s2	0.250 (710)	0.188	0.217
average s3	0.189 (704)	0.145	0.165
average s4	0.195 (674)	0.167	0.180

Conclusion

- Meta information: paraphrasing by NPLM
- NPLM captures the semantically and syntactically similar words in a way that a latent word depends on the context.
- Plain paraphrasing: lost 0.39 BLEU points absolute compared to the standard confusion network-based system combination (Probably because of noise).
- Paraphrasing with assessment: lost 0.19 BLEU points absolute.

Acknowledgement

Thank you for your attention.

- This research is supported by the the 7th Framework Programme and the ICT Policy Support Programme of the European Commission through the T4ME project (Grant agreement No. 249119).
- This research is supported by the Science Foundation Ireland (Grant 07/CE/I1142) as part of the Centre for Next Generation Localisation at Dublin City University.