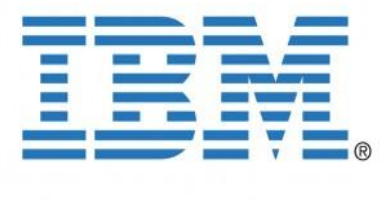
Goodness: A Method for Measuring Machine Translation Confidence

Nguyen Bach

Carnegie Mellon University

Fei Huang and Yaser Al-Onaizan

IBM T.J. Watson Research Center



Source

أنت مختلف تماماً عن زيد وعمرو فلا تحشر نفسك في سرداب التقليد والمحاكاة والذوبان

MT output

you totally different from zaid amr , and not to deprive yourself in a basement of imitation and assimilation .

We predict and visualize

you **totally** different from **zaid amr , and** **not to deprive yourself** **in a basement of imitation and** assimilation .

Human correction

you are quite different from zaid and amr , so do not cram yourself in the tunnel of simulation , imitation and assimilation .

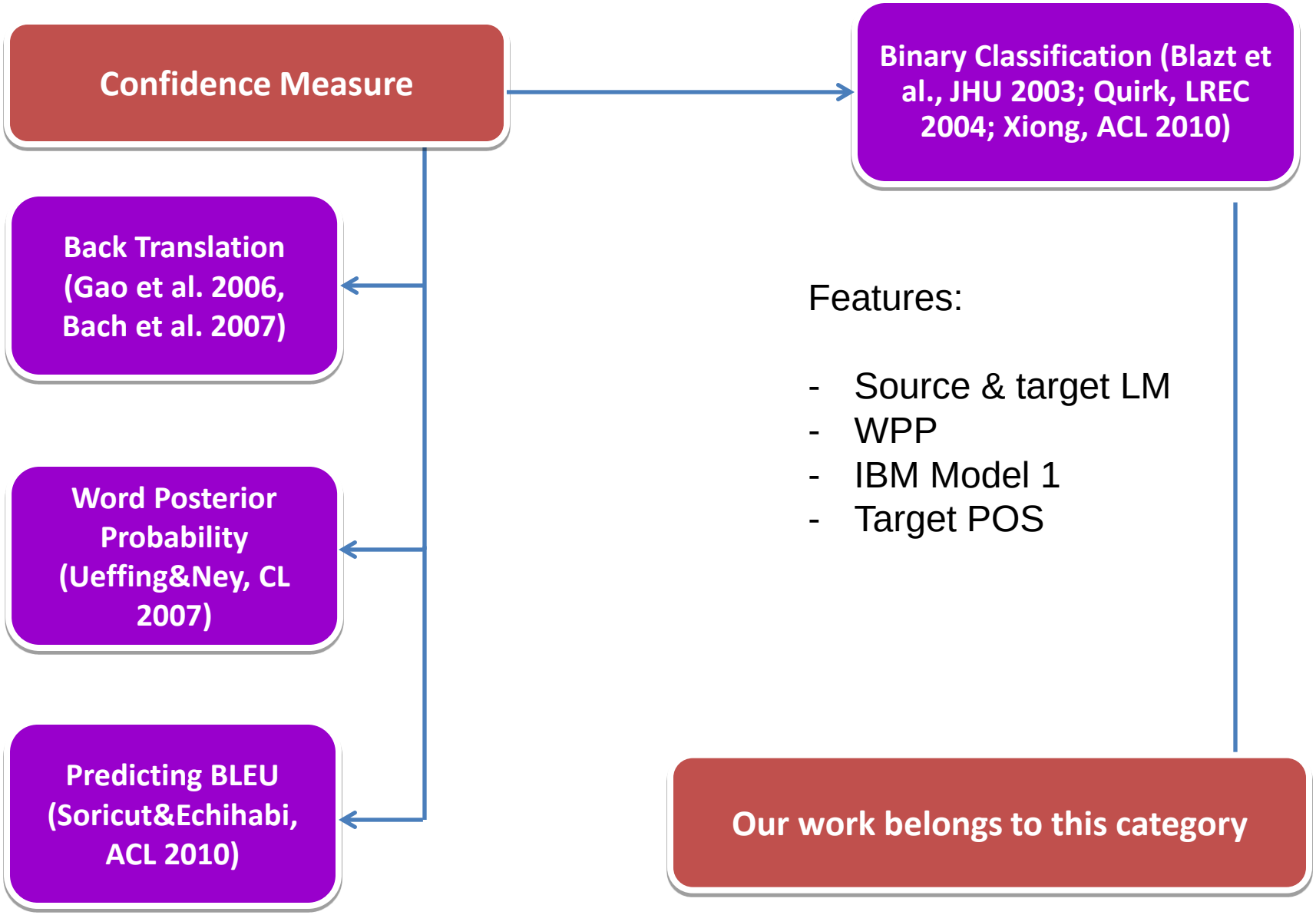
Why does it matter?

- Current MT systems don't have efficient ways to predict machine translation confidence.

- Applications
  - Post-editing
  - Commercial practice for end-users
  - Down-stream processing: QA, IE
  - MT engine: reranking

Outline

- Introduction
- Predicting
  - Review
  - Our models
  - Feature sets
- Experiments
- N-best list reranking
- Visualizing


Confidence Measure
Binary Classification (Blazt et al., JHU 2003; Quirk, LREC 2004; Xiong, ACL 2010)
Back Translation (Gao et al. 2006, Bach et al. 2007)
Word Posterior Probability (Ueffing&Ney, CL 2007)
Predicting BLEU (Soricut&Echihabi, ACL 2010)
Features:
- Source & target LM
- WPP
- IBM Model 1
- Target POS
Our work belongs to this category

Can we do a better job?

- Can we use a rich feature set such as dependency structures, source-side information, and alignment context to improve error prediction performance?
- Can we predict translation error types?
- Do confidence measures help the MT system to select a better translation?
- How confidence score can be presented to improve end-user perception?

Confidence Measure Models

- Confidence estimation ≈ sequential labelling task
  - Word Sequence = MT output
- Two classifiers
  - Binary: Good/Bad
  - Multi classes: Insertion/Substitution/Shift/Good.
- Supervised training
- Word -> Sentence confidence

Word-level Model

- A feature-rich classifier for a given word

$$score(f, y) = \sum_i f_i w_i^y$$

where *f* is its feature vector, *y* is the label, w^y is its feature weight vector

- Discriminatively trained by MIRA

Sentence-level Model

- Given a sentence S, use word-level models to obtain labels for each words

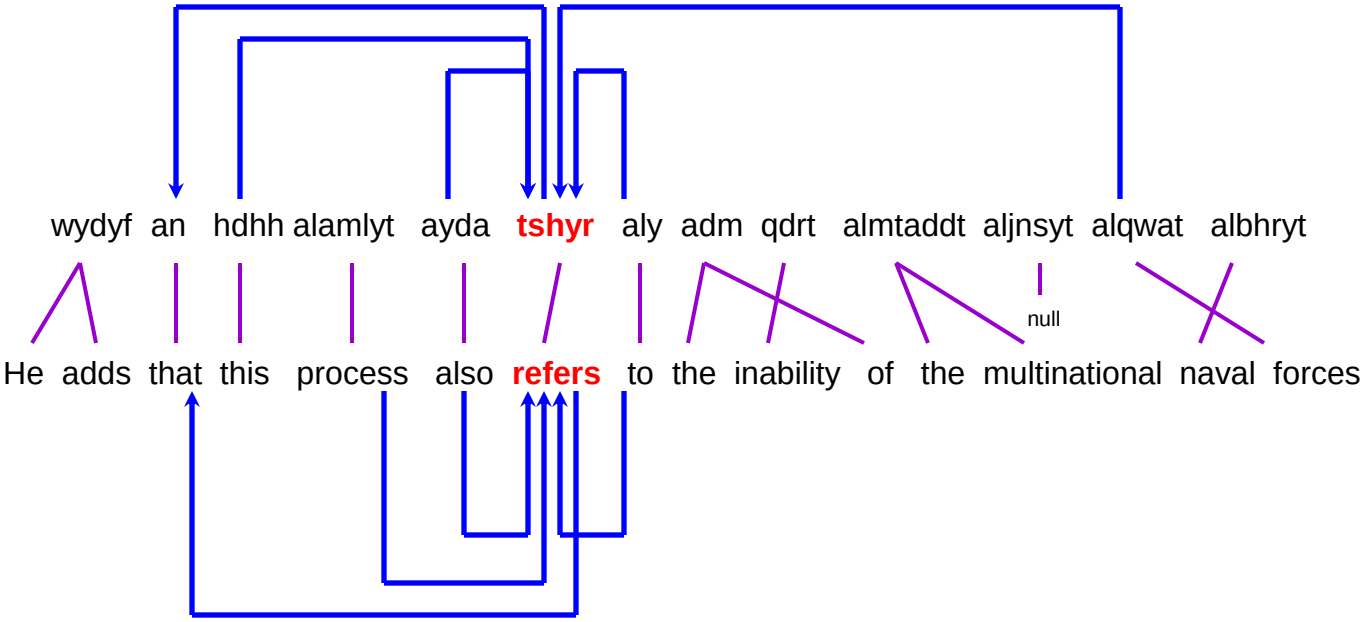
- Use forward and backward values to compute marginal probability of Good label

$$p(y_i = Good|S) = \frac{\alpha(y_i|S)\beta(y_i|S)}{\sum_j \alpha(y_j|S)\beta(y_j|S)}$$

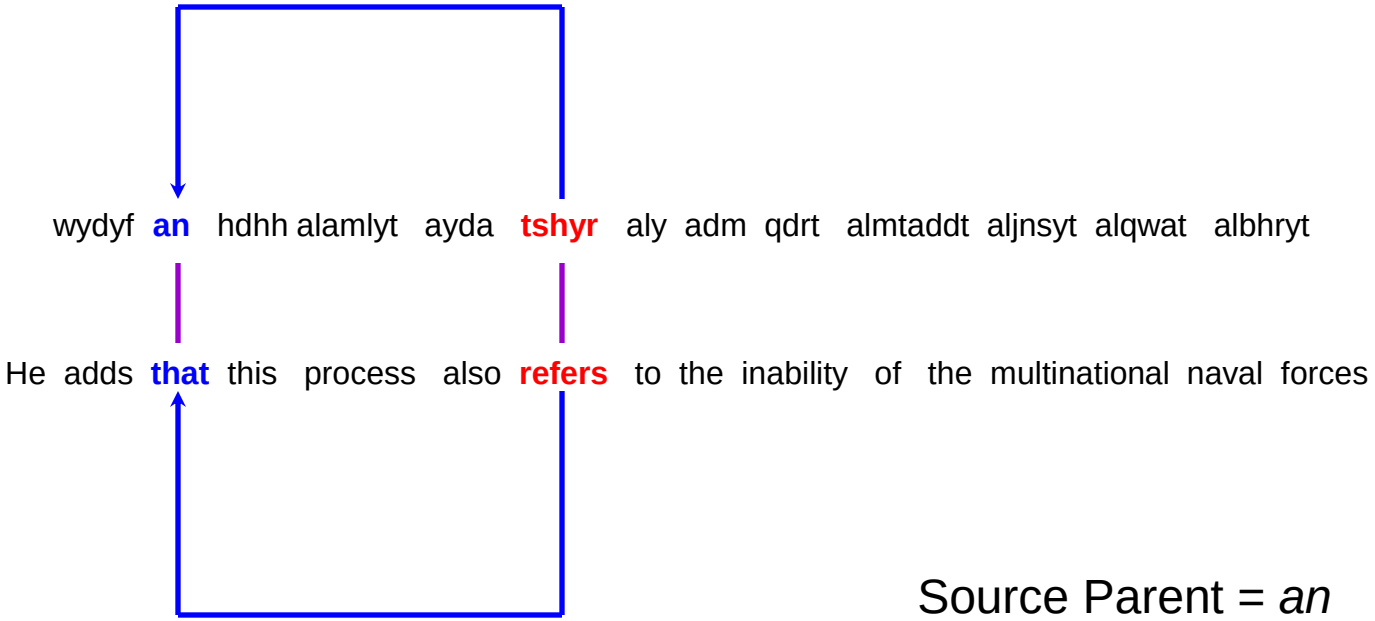
- Normalize

$$goodness(S) = \frac{\sum_{i=1}^{k} p(y_i{=}Good|S)}{k}$$

Type 1: Source & Target Dependency Structures



Type 1: Source & Target Dependency Structures



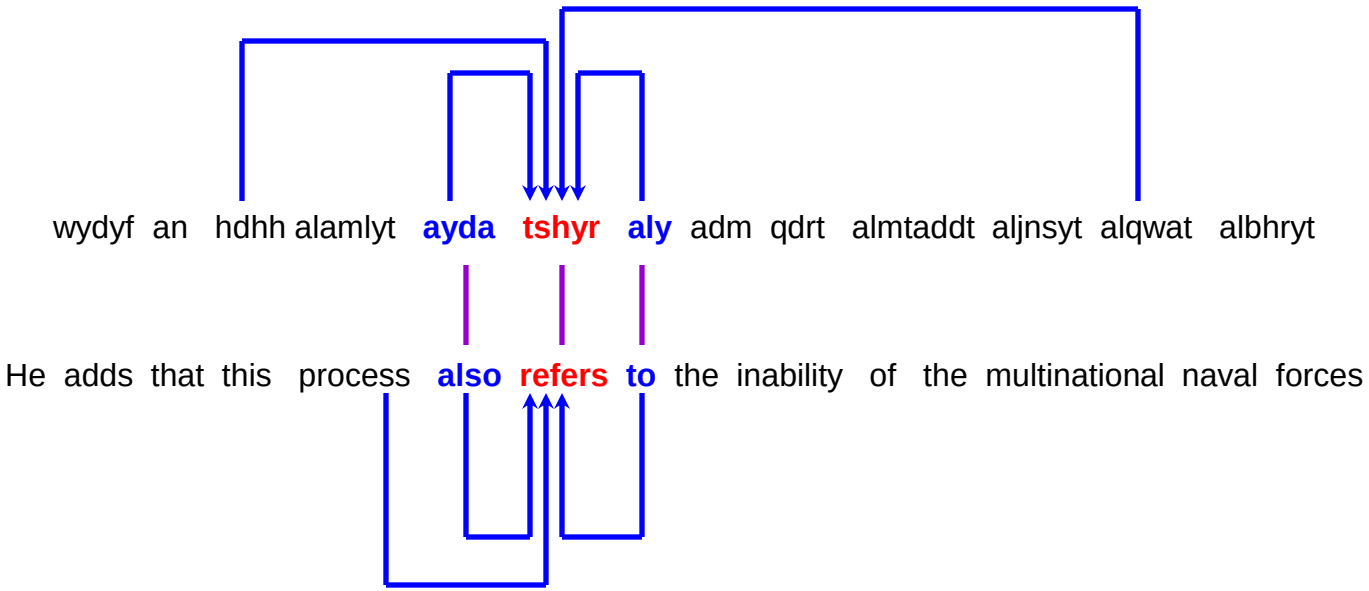
Child-Parent Agreement

Source Parent = *an*

Target Parent = *that*

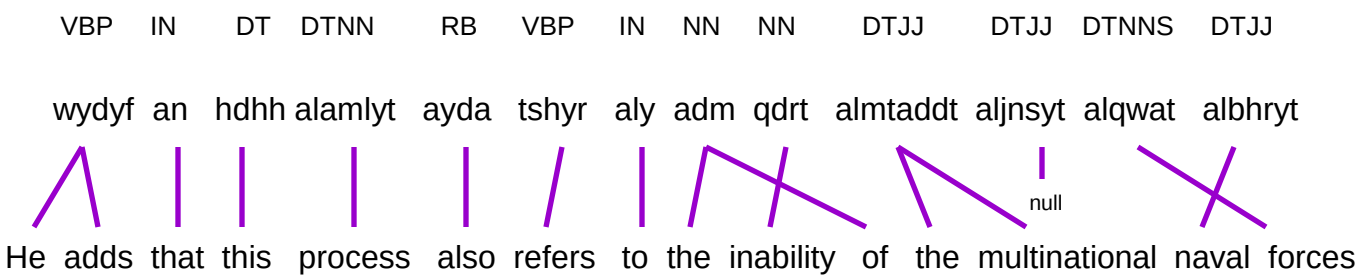
Are parent aligned: Yes

Type 1: Source & Target Dependency Structures



Children Agreement: 2

Type 2: Source POS and Phrases



Type 2: Source POS & Phrases

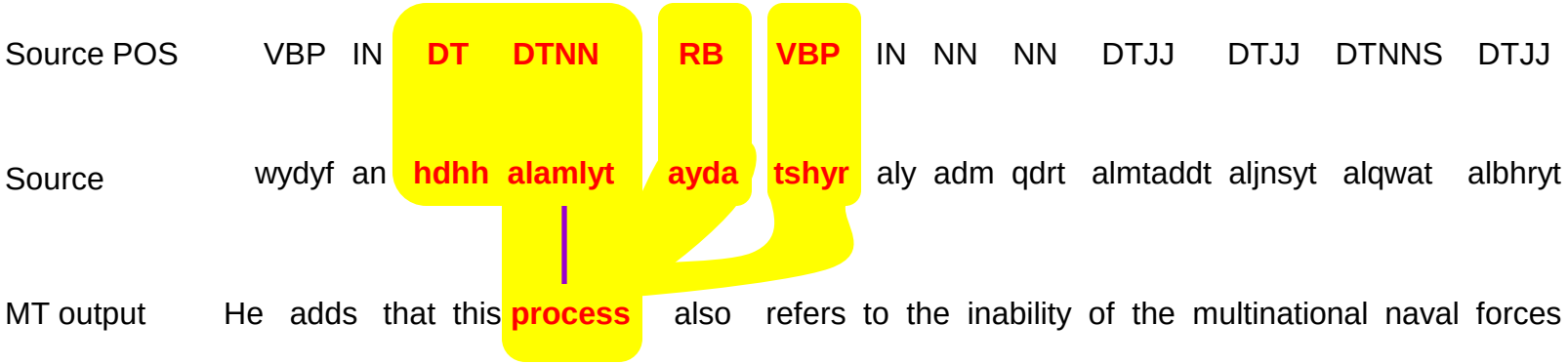
Source POS	VBP	IN	DT	DTNN	RB	VBP	IN	NN	NN	DTJJ	DTJJ	DTNNS	DTJJ
Source	wydyf	an	**hdhh**	**alamlyt**	ayda	tshyr	aly	adm	qdrt	almtaddt	aljnsyt	alqwat	albhryt
MT output	He	adds	that	this	**process**	also	refers	to	the	inability	of	the	multinational	naval	forces

Type 2: Source POS and Phrases

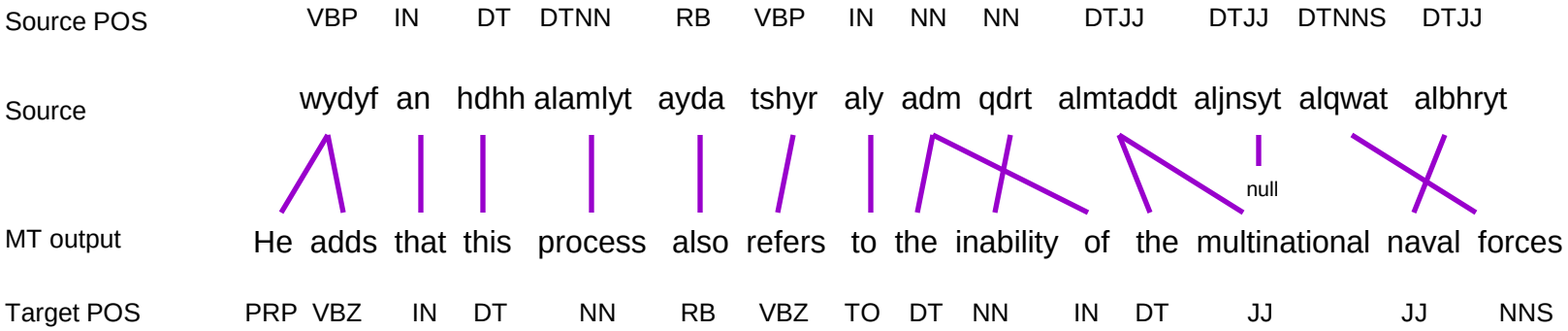
Source POS	VBP	IN	**DT**	**DTNN**	RB	VBP	IN	NN	NN	DTJJ	DTJJ	DTNNS	DTJJ	
Source	wydyf	an	**hdhh**	**alamlyt**	ayda	tshyr	aly	adm	qdrt	almtaddt	aljnsyt	alqwat	albhryt	
MT output	He	adds	that	this	**process**	also	refers	to	the	inability	of	the	multinational	naval	forces

f_{125}(target-word = "*process*") = { 1 if source-POS-sequence = "*DT DTNN*"; 0 otherwise

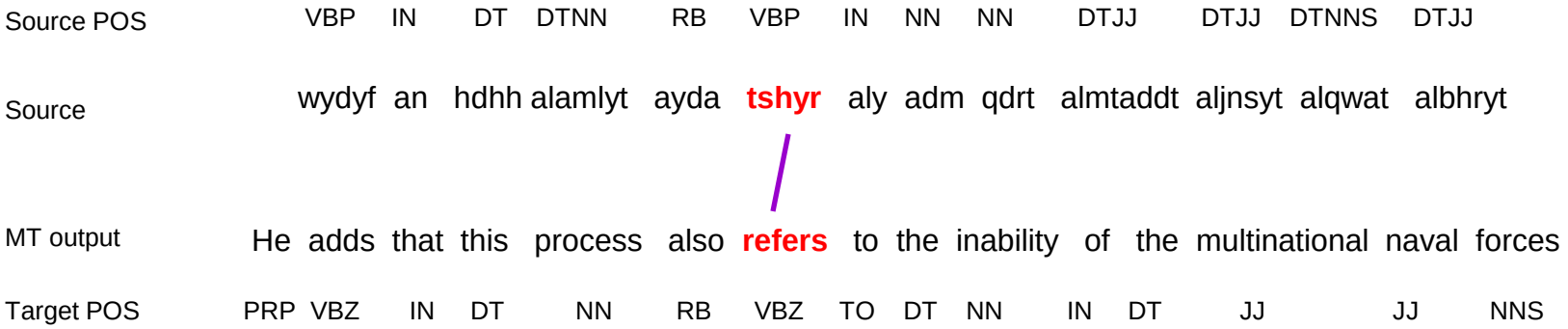
Type 2: Source POS and Phrases



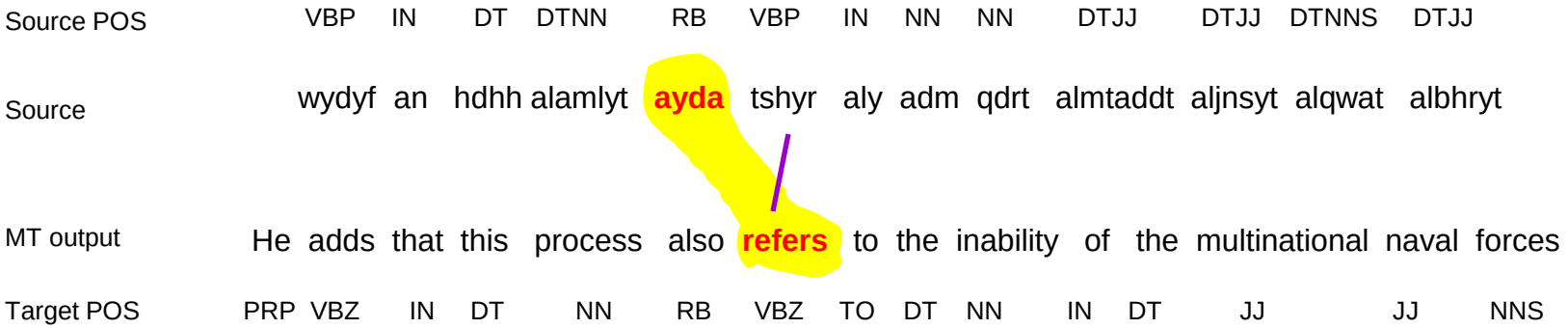
Type 3: Alignment Context



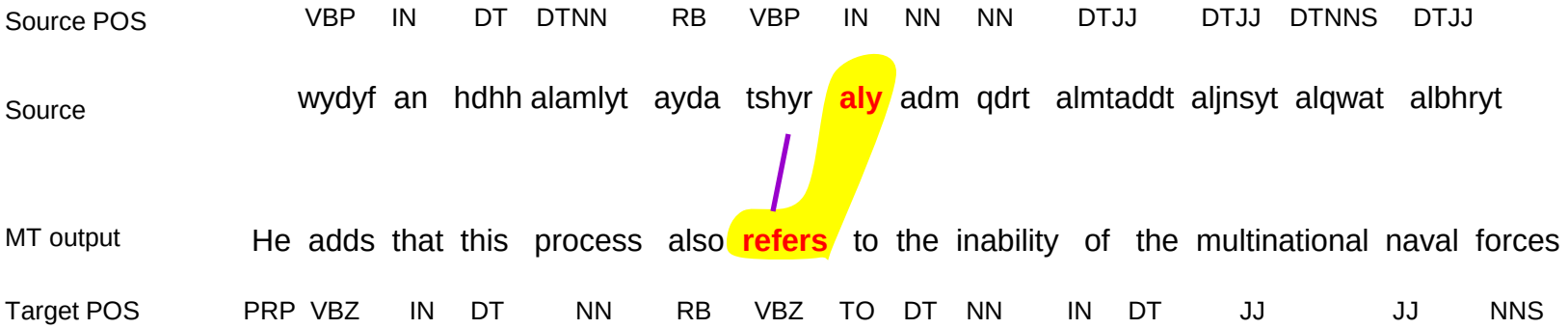
Type 3: Alignment Context



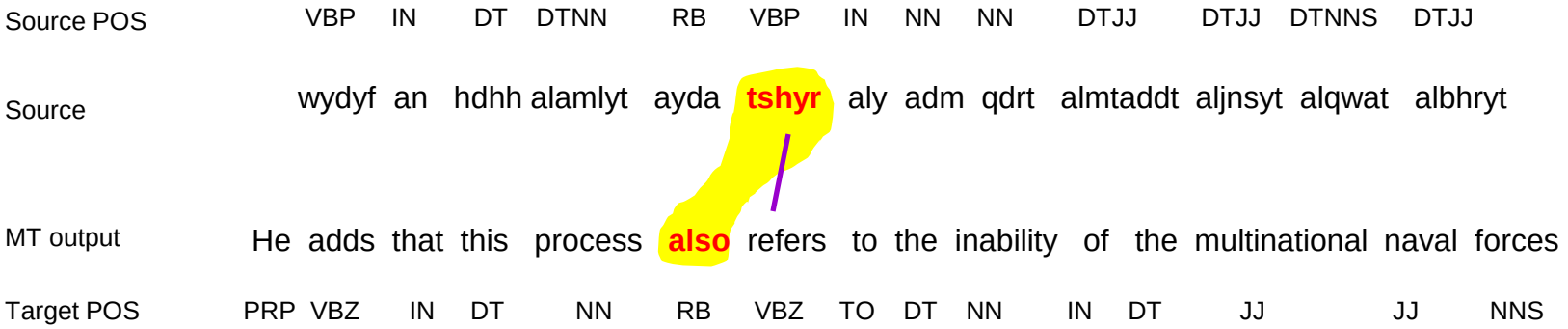
Type 3: Alignment Context



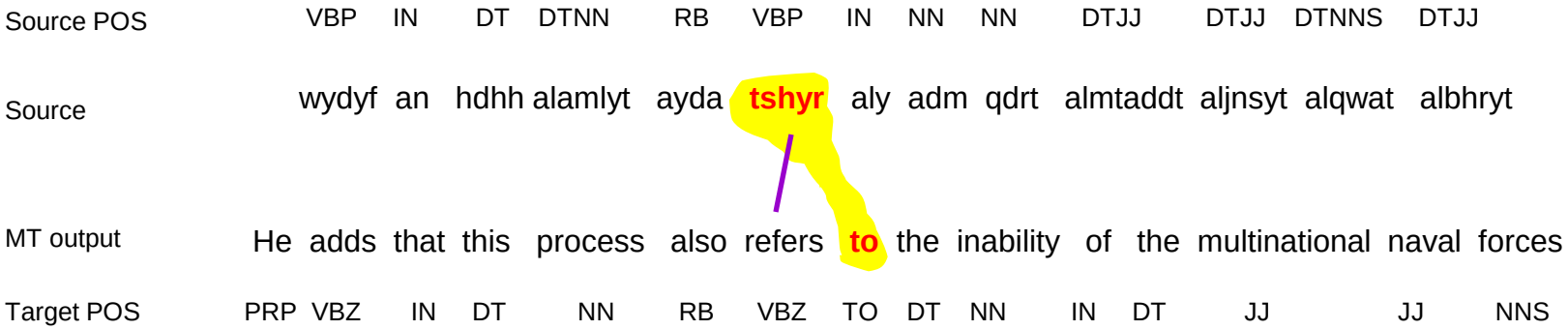
Type 3: Alignment Context



Type 3: Alignment Context

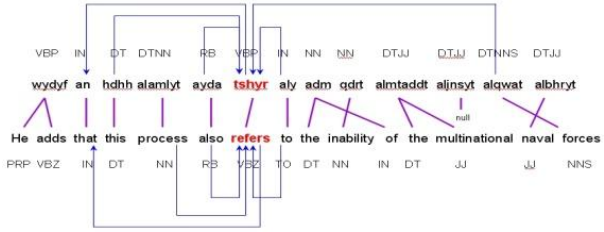


Type 3: Alignment Context

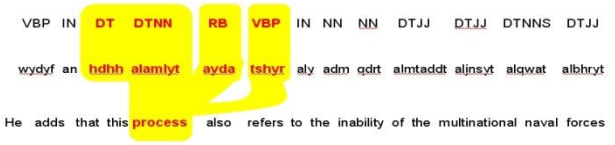


Recap

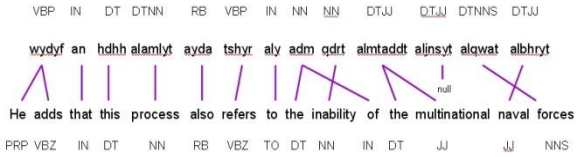
- **Source & Target Dependency**



- **Source POS and Phrases**



- **Alignment Context**



Experiment Setup

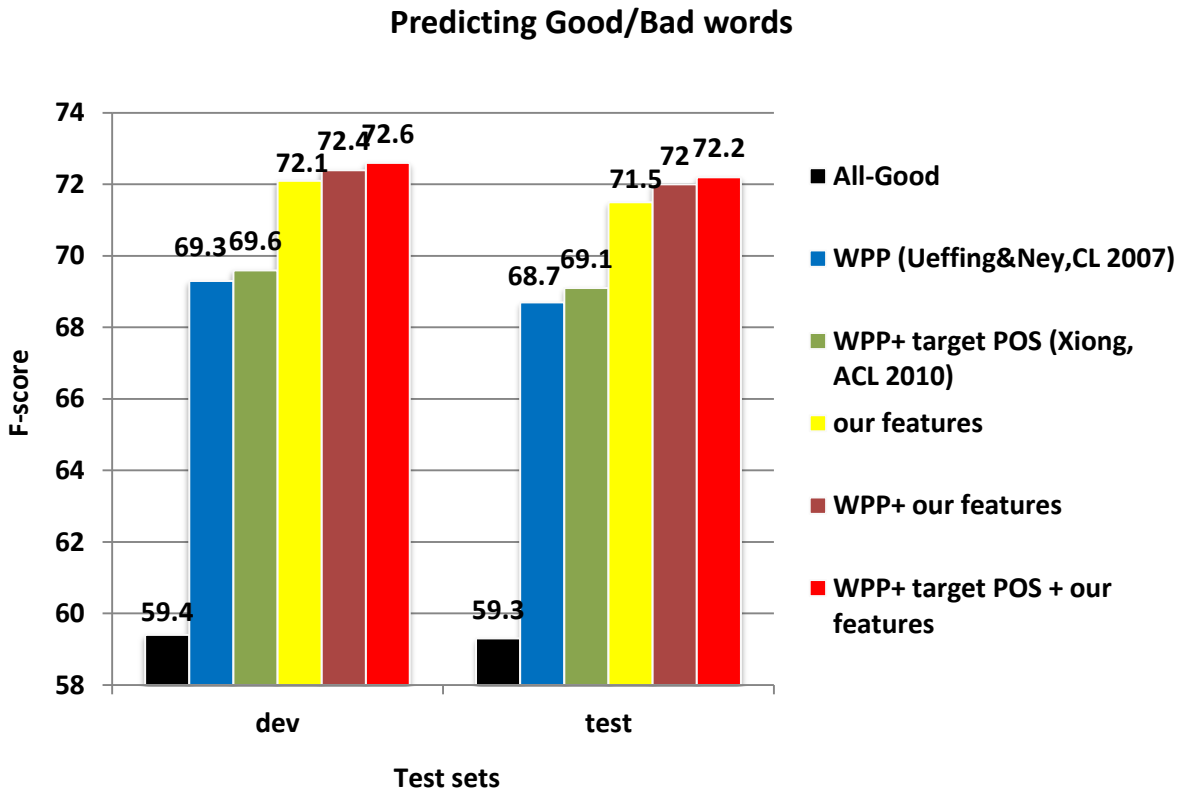
Source	أنت مختلف تماماً عن زيد وعمرو فلا تحشر نفسك في سرداب التقليد والمحاكاة والذوبان
MT output	you totally different from zaid amr , and not to deprive yourself in a basement of imitation and assimilation .
Human correction	you are quite different from zaid and amr , so do not cram yourself in the tunnel of simulation , imitation and assimilation .

Experiment Setup

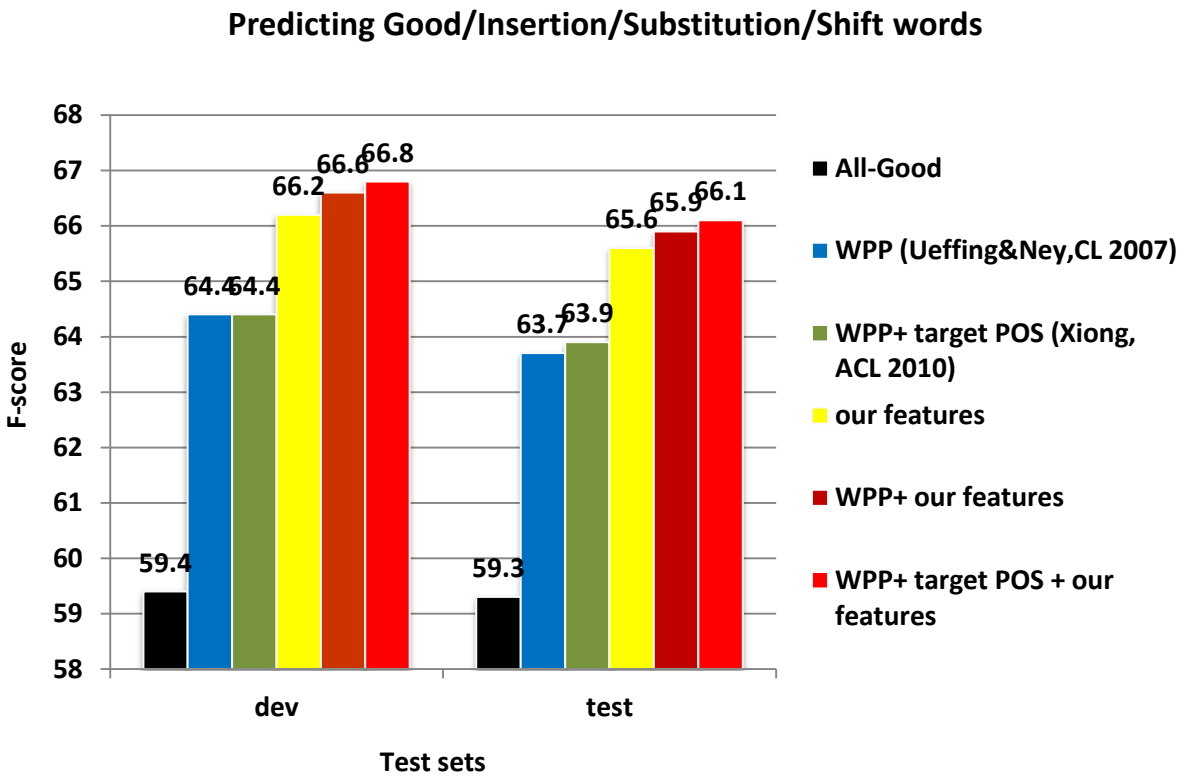
- Training data
  - Human correction Arabic-English system: 72K sentences with 2.2M words; mix newswire and weblog
- Dev and Unseen test set
  - Dev: 2,707 sentences, 80K words
  - Unseen: 2,707 sentences, 79K words
- WPP: implement WPP algorithm in (Ueffing&Ney, 2007)
- WPP+target-POS: similar feature set as in (Xiong, ACL 2010)
- Learner: MIRA, 100 iterations, hyper-param = 5, cut-off = 1
- Evaluation metric: F-score

How does the proposed method compare with previous work?

Performance of binary classifiers



Performance of 4-class classifiers

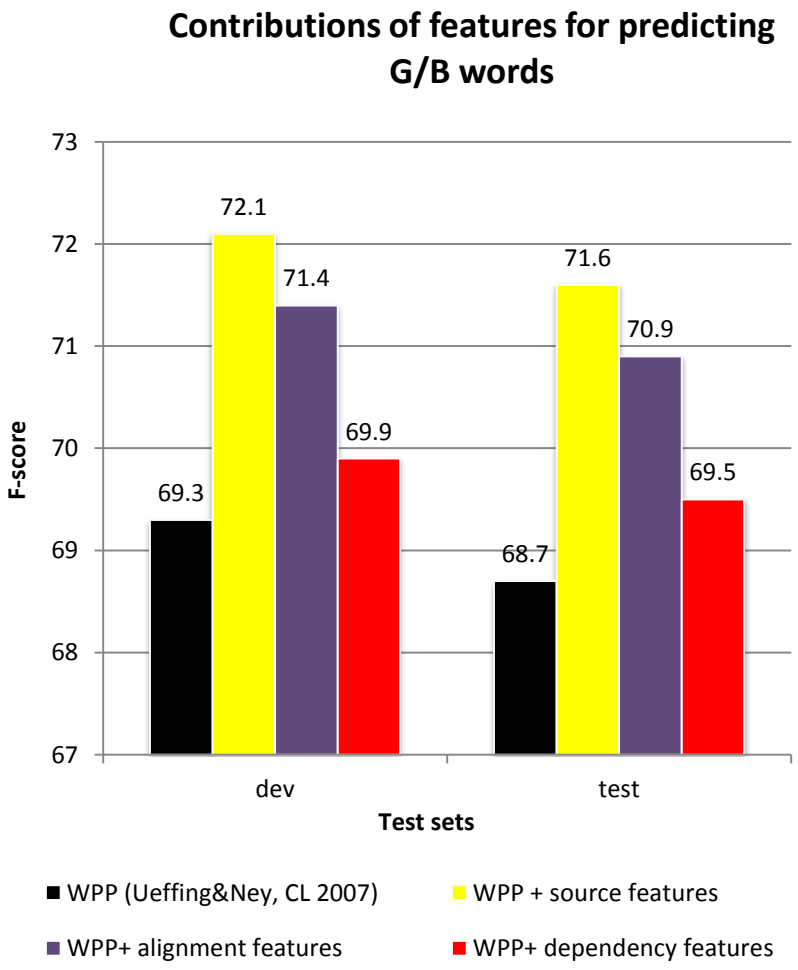


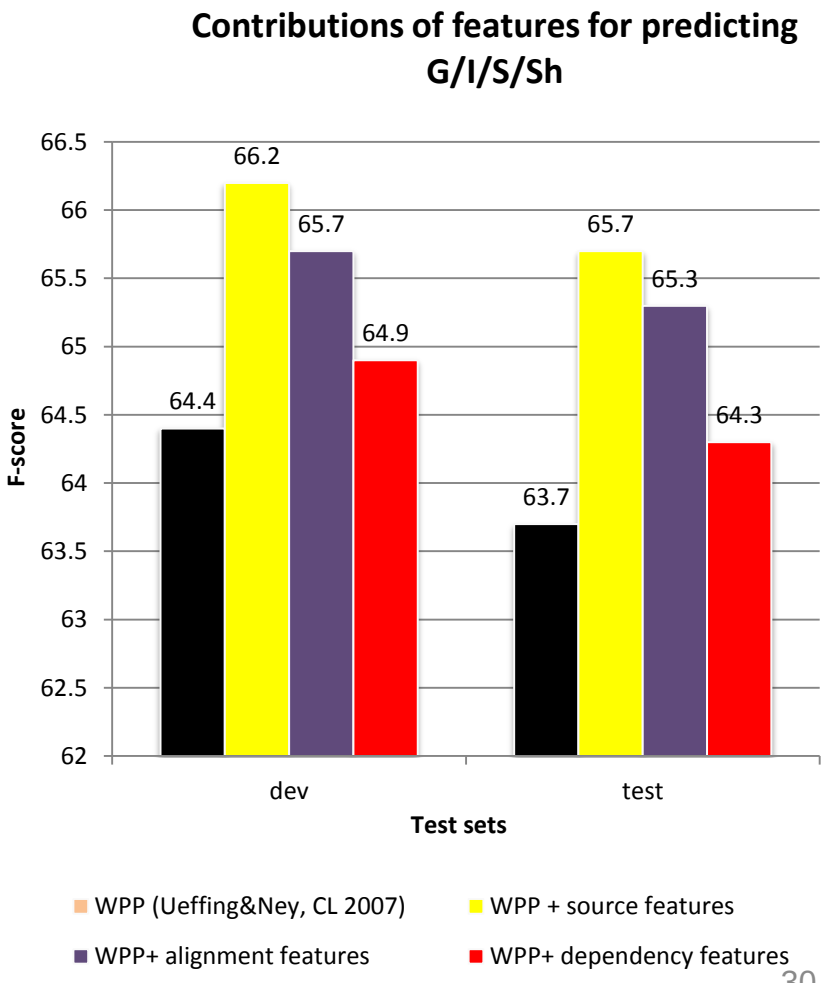
- Improve from WPP (Ueffing & Ney, CL 2007) and WPP+targetPOS (Xiong et al., ACL 2010)

- Our features alone already outperformed previous work

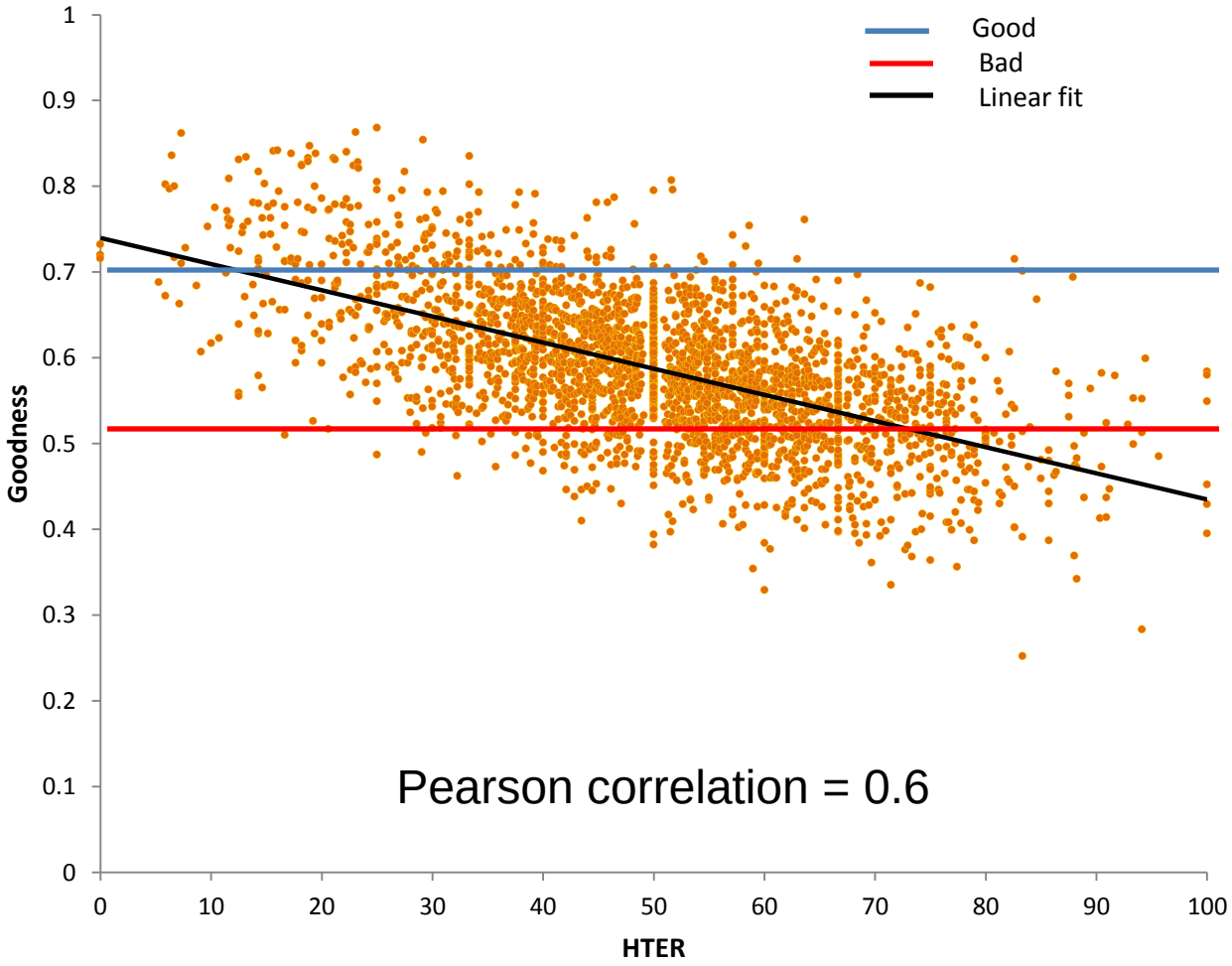
What are the contributions of individual feature sets?

Where are improvements coming from?





Goodness vs. HTER



Outline

- Introduction
- Predicting
  - Review
  - Our models
  - Feature sets
- Experiments
- N-best list reranking
- Visualizing

N-best list reranking

- Reranking n-best list by *goodness*

	Dev		Test	
	TER	BLEU	TER	BLEU
Baseline	49.9	31	50.2	30.6
2-best	49.5	31.4	49.9	30.8
5-best	49.2	31.4	49.6	30.8
50-best	49.1	30.9	49.4	30.5
100-best	49	30.9	49.3	30.5

N-best list reranking

- Reranking n-best list by *goodness*

	Dev		Test	
	TER	BLEU	TER	BLEU
Baseline	49.9	31	50.2	30.6
2-best	49.5	31.4	49.9	30.8
5-best	49.2	31.4	49.6	30.8
50-best	49.1	30.9	49.4	30.5
100-best	49	30.9	49.3	30.5

N-best list reranking

- Reranking n-best list by *goodness*

	Dev		Test	
	TER	BLEU	TER	BLEU
Baseline	49.9	31	50.2	30.6
2-best	49.5	31.4	49.9	30.8
5-best	49.2	31.4	49.6	30.8
50-best	49.1	30.9	49.4	30.5
100-best	49	30.9	49.3	30.5

Outline

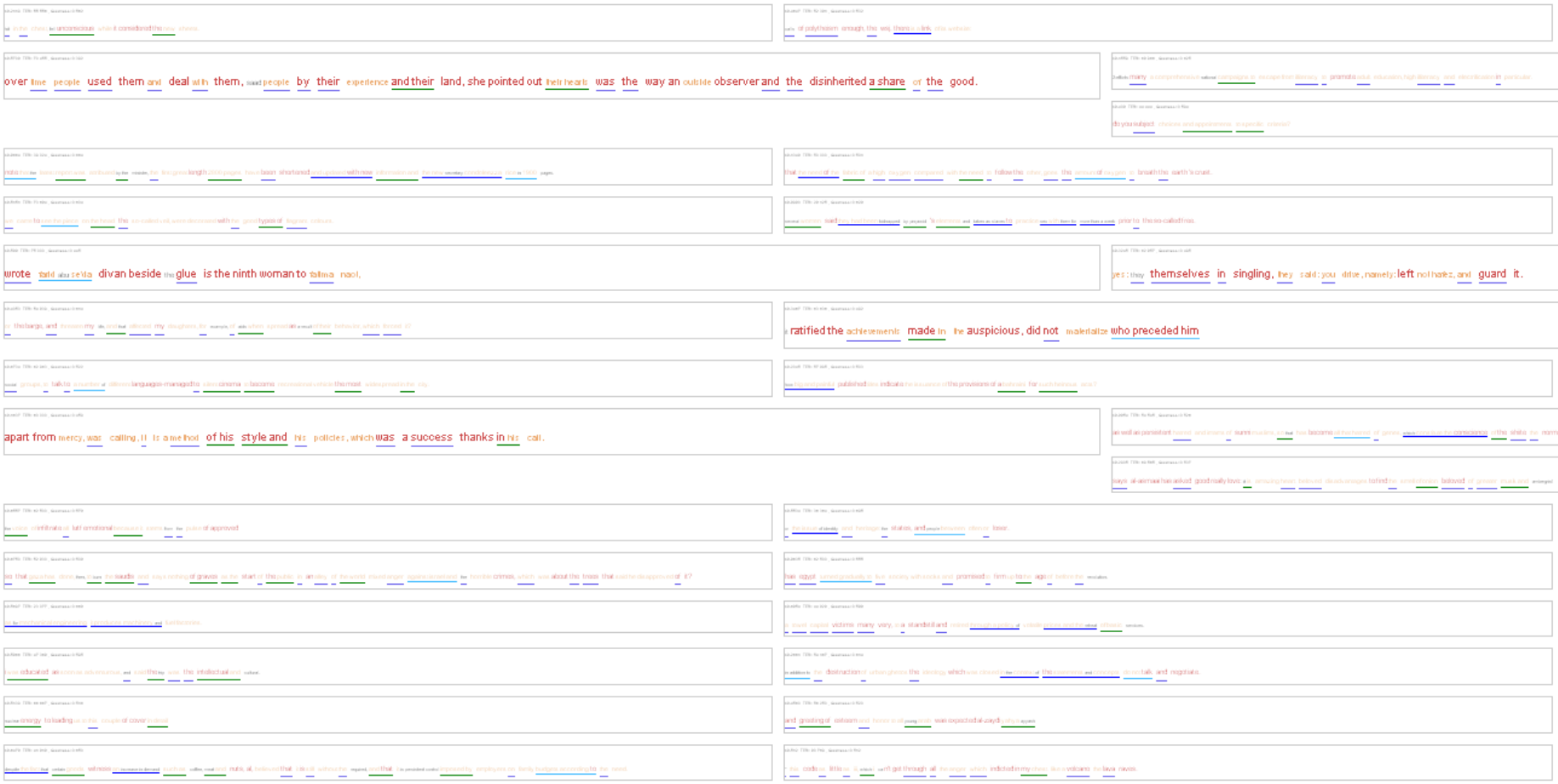
- Introduction
- Predicting
  - Review
  - Our models
  - Feature sets
- Experiments
- N-best list reranking
- Visualizing

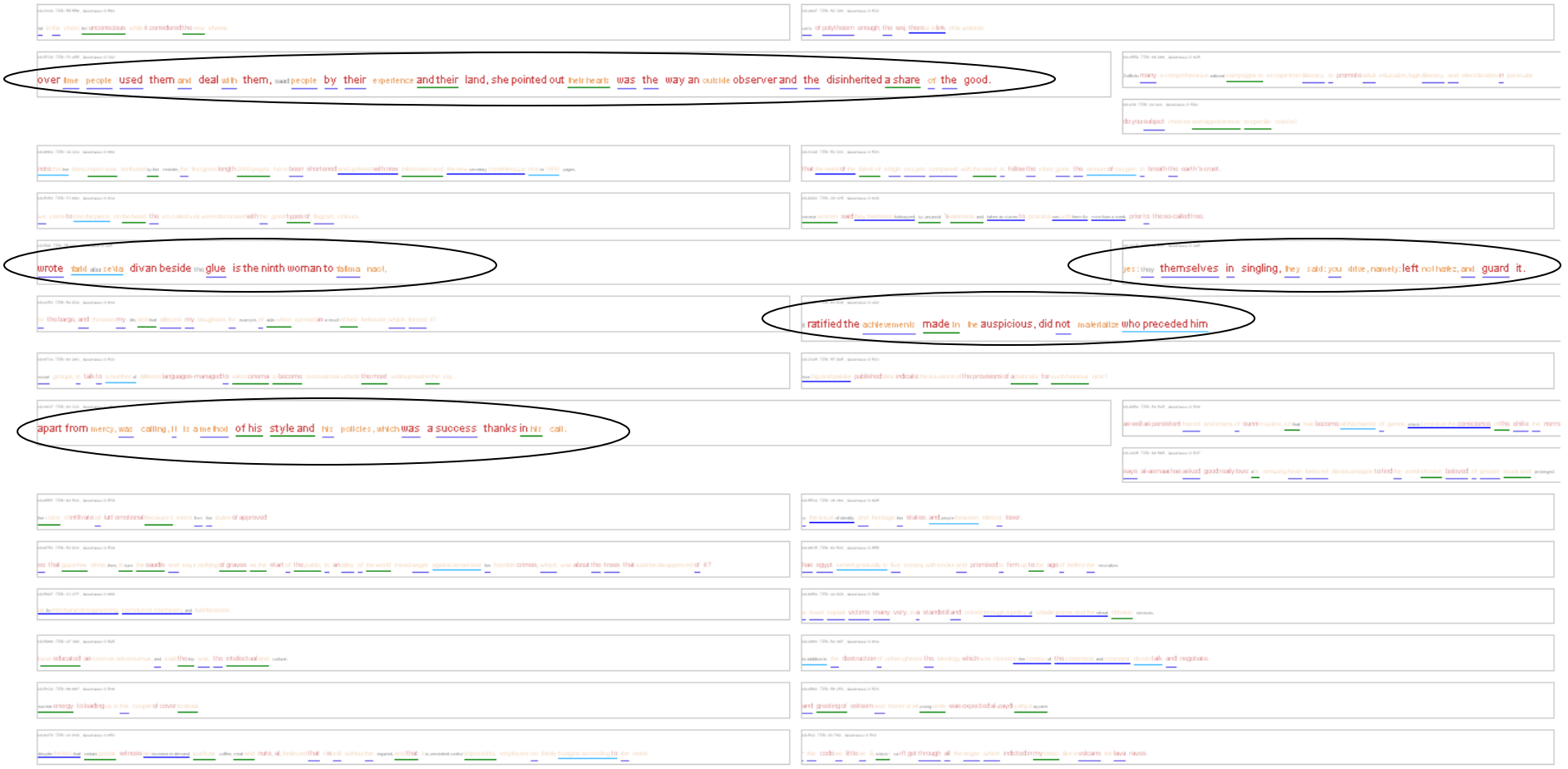
Designing Objectives

- Minimize post editing efforts in word-level and sentence-level
- Simple and Intuitive
  - Inspired by Wordle ©IBM

Designing Objectives

- Font size
  - Big: bad
  - Small: good
  - Medium: decent
- Color
  - **Red: bad**
  - Black: good
  - Orange: decent
- Threshold
  - Good word > 0.8; Bad word < 0.45
  - Good sentence > 0.7; Bad sentence < 0.5


over time people used them and deal with them, said people by their experience and their land, she pointed out their hearts was the way an outside observer and the disinherited a share of the good.
wrote farid abu salma divan beside the glue is the ninth woman to salma naol,
apart from mercy, was calling, it is a method of his style and his policies, which was a success thanks in his call.



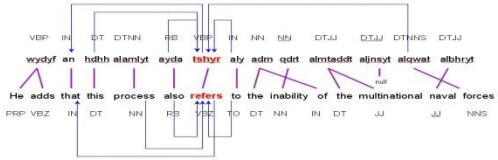
Bad translations: editing these sentences first

Source	واظهر الاستطلاع ايضا ان معظم المشاركين في الدول النامية مستعدون لادخال تغييرات نوعية على نمط حياتهم في سبيل خفض تأثيرات التغير المناخي .
MT output	the poll also showed that most of the participants in the developing countries are ready to introduce qualitative changes in the pattern of their lives for the sake of reducing the effects of climate change.
We predict and visualize	the poll also showed that most of the participants in the developing countries are ready to introduce **qualitative** changes in the pattern of their lives for the sake of reducing the effects of climate change.
Human correction	the survey also showed that most of the participants in developing countries are ready to introduce changes to the quality of their lifestyle in order to reduce the effects of climate change .

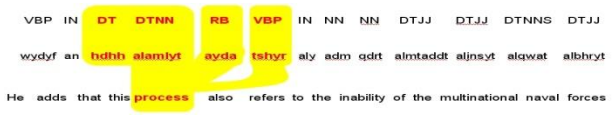
Source	لعل العنصرية من اوسأ امراض البشرية ومن اصعبها علاجا ولا ينجو منها الا من استعان بحبل من الله
MT output	perhaps racism of heads of human diseases and the most difficult treatment, survives, only used rope,
We predict and visualize	perhaps racism **of** heads of human diseases and the most difficult treatment, **survives, only used rope,**
Human correction	racism is but the worst human disease and the hardest to recover ; just he who is faithful to allah is able to recover .

Contributions

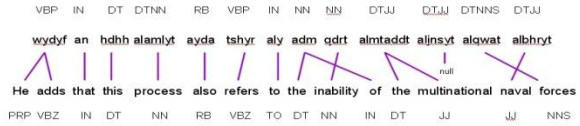
- Source & Target Dependency Structures



- Source POS and Phrases



- Word Alignment Context



- Experimental results
  - Our features outperformed features described in the previous work
- Improve MT quality through n-best list reranking
- Visualizing confidence in word and sentence level

Thanks!