

Error Detection for SMT Using Linguistic Features

Deyi Xiong, Min Zhang, and Haizhou Li
Human Language Technology
Institute for Infocomm Research



What's error detection: an example

SRC: 国际足联将严惩足球场上的欺骗行为

What's error detection: an example

SRC: 国际足联将严惩足球场上的欺骗行为

REF: FIFA will severely punish all cheating acts in the football field

What's error detection: an example

SRC: 国际足联将严惩足球场上的欺骗行为

REF: FIFA will severely punish all cheating acts in the football field

SYS: FIFA will severely punish fraud on the pitch

What's error detection: an example

SRC: 国际足联将严惩足球场上的欺骗行为

REF: FIFA will severely punish all cheating acts in the football field

SYS: FIFA will severely punish fraud on the pitch



correct



incorrect

Applications

- Post-editing
- Interactive machine translation
- Can help SMT too
 - Rescoring
 - Regenerating
 - System combination

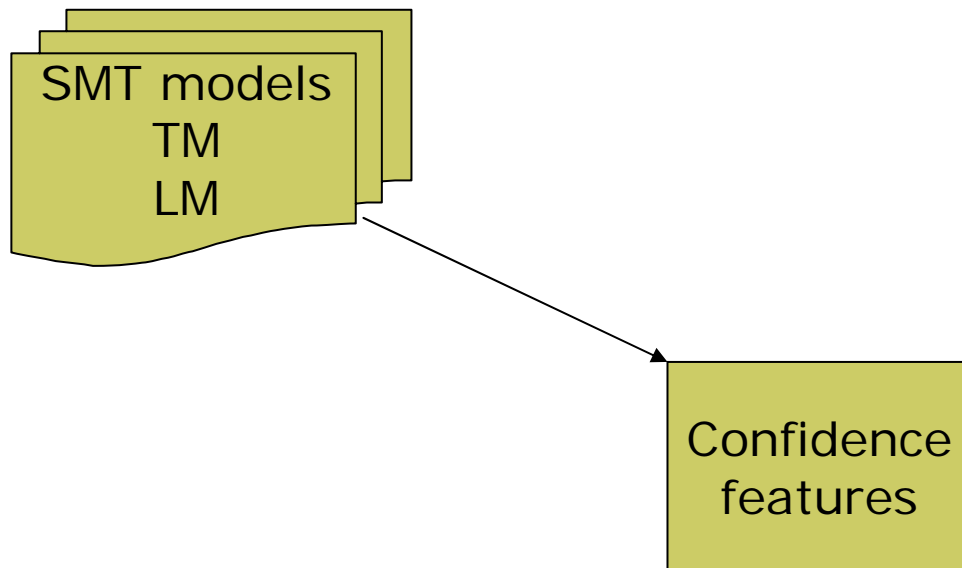
Previous work

- **Confidence estimation** (Blatz et al., 2003; Ueffing & Ney, 2007; Sanchis et al., 2007)
 - Calculate the confidence at which a word is correct

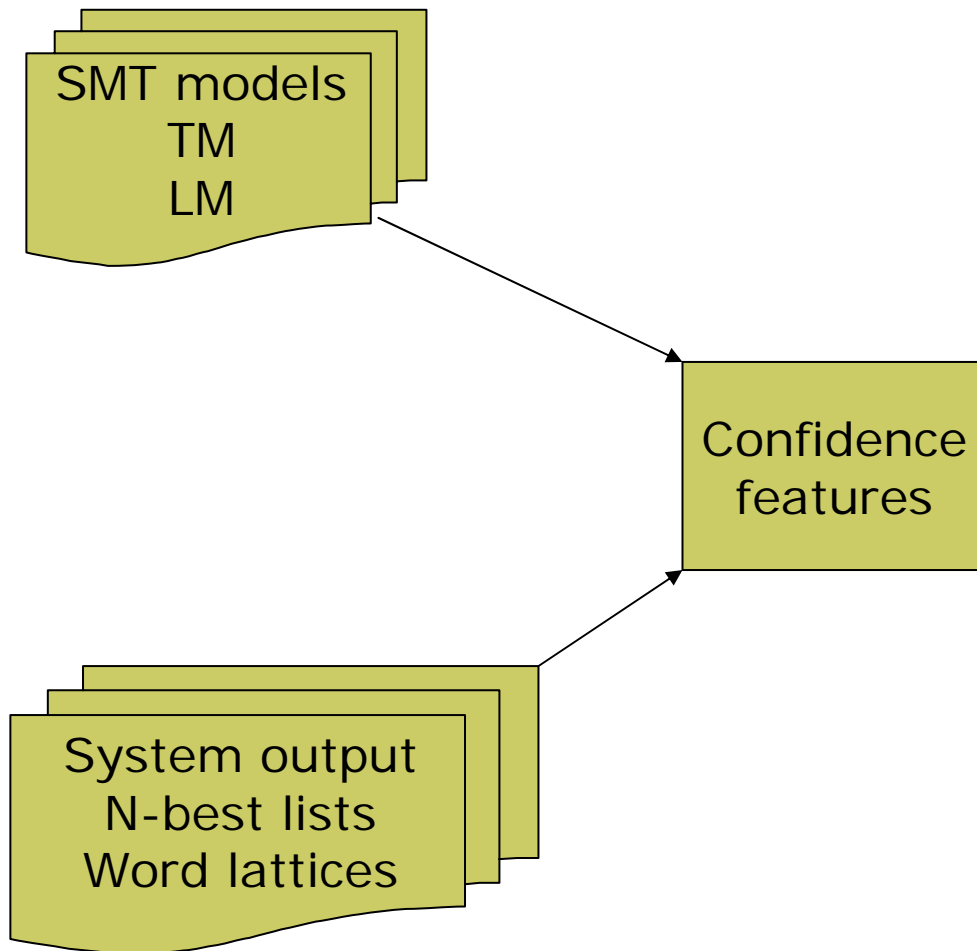
Word confidence estimation

Confidence
features

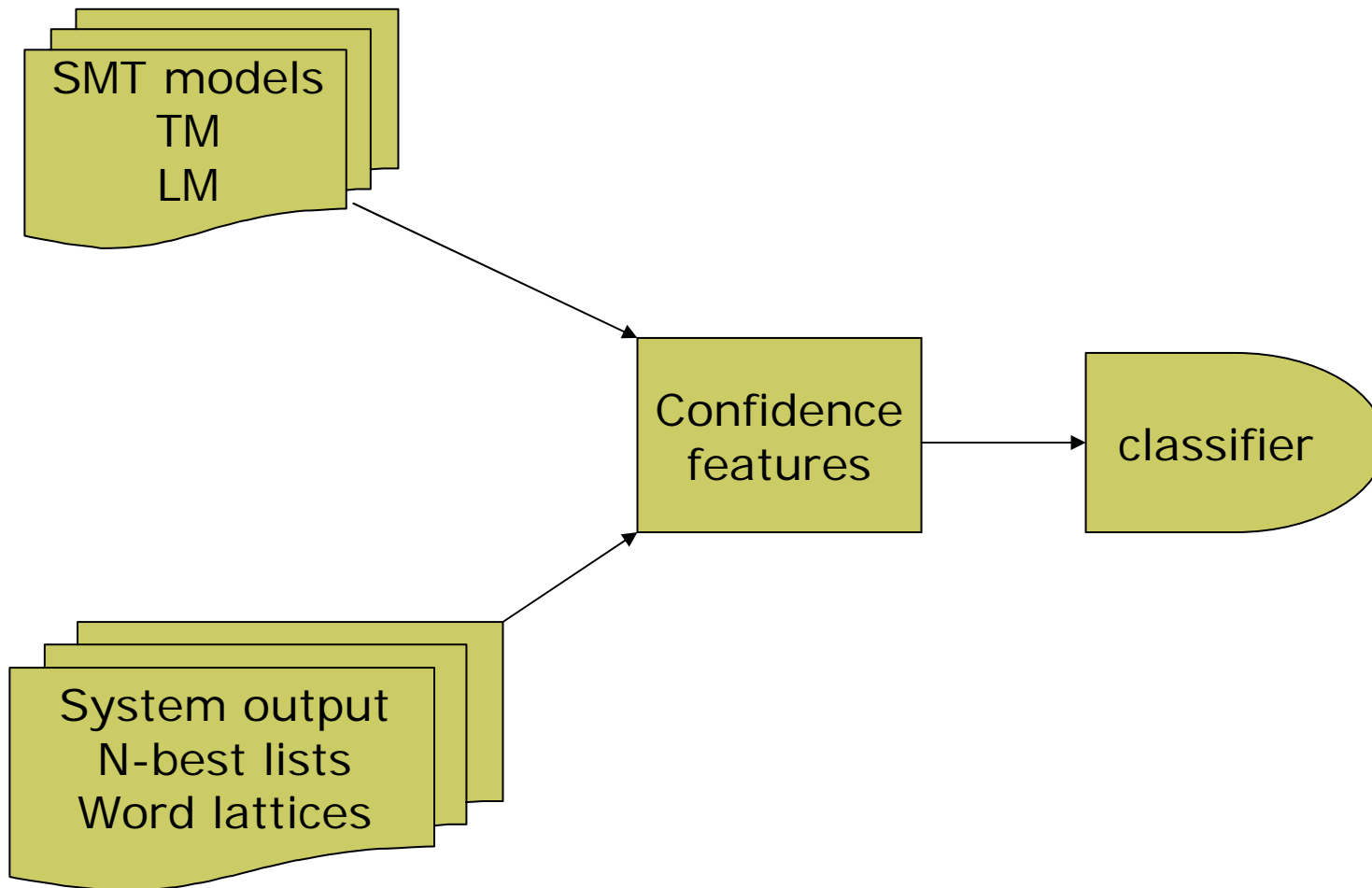
Word confidence estimation



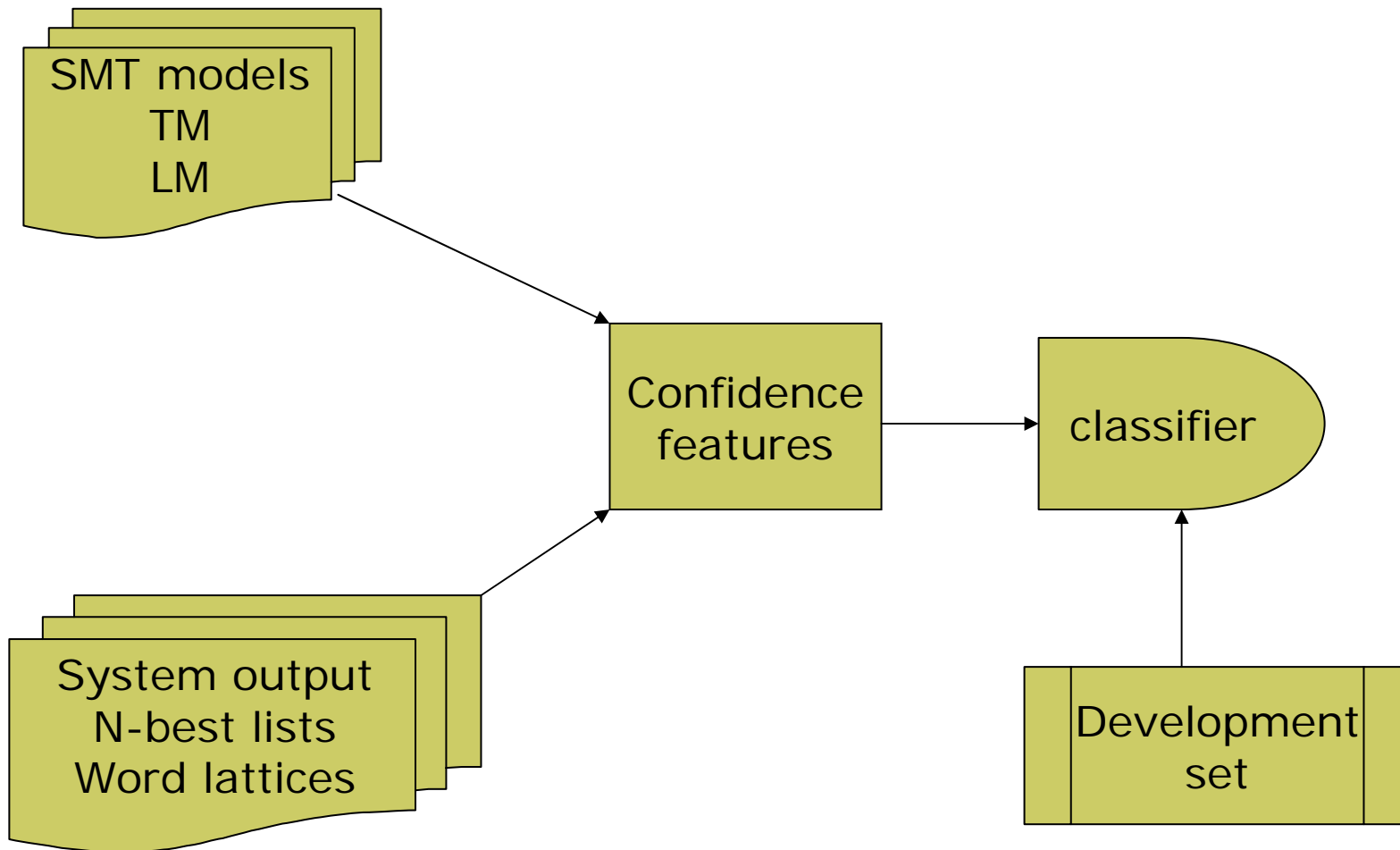
Word confidence estimation



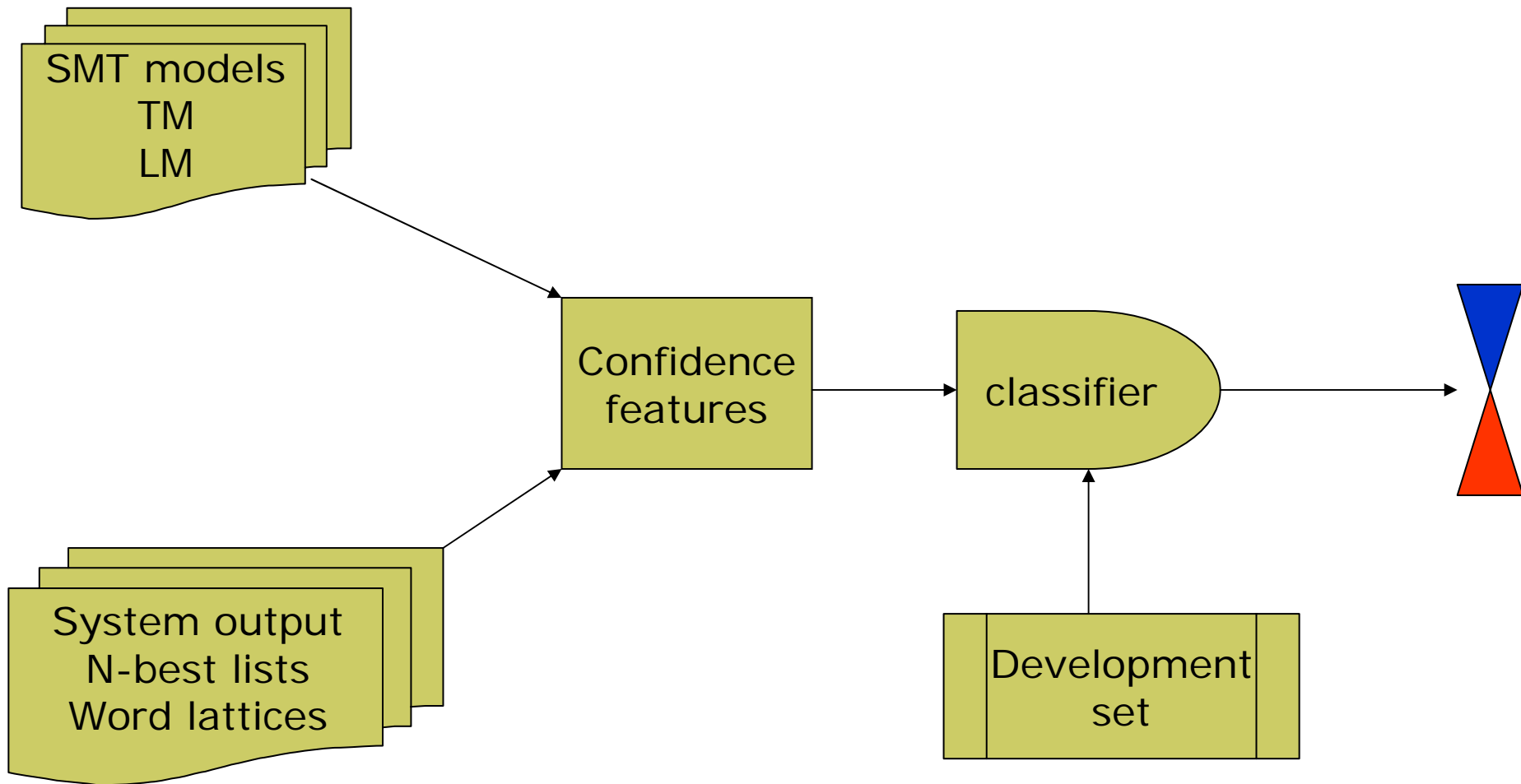
Word confidence estimation



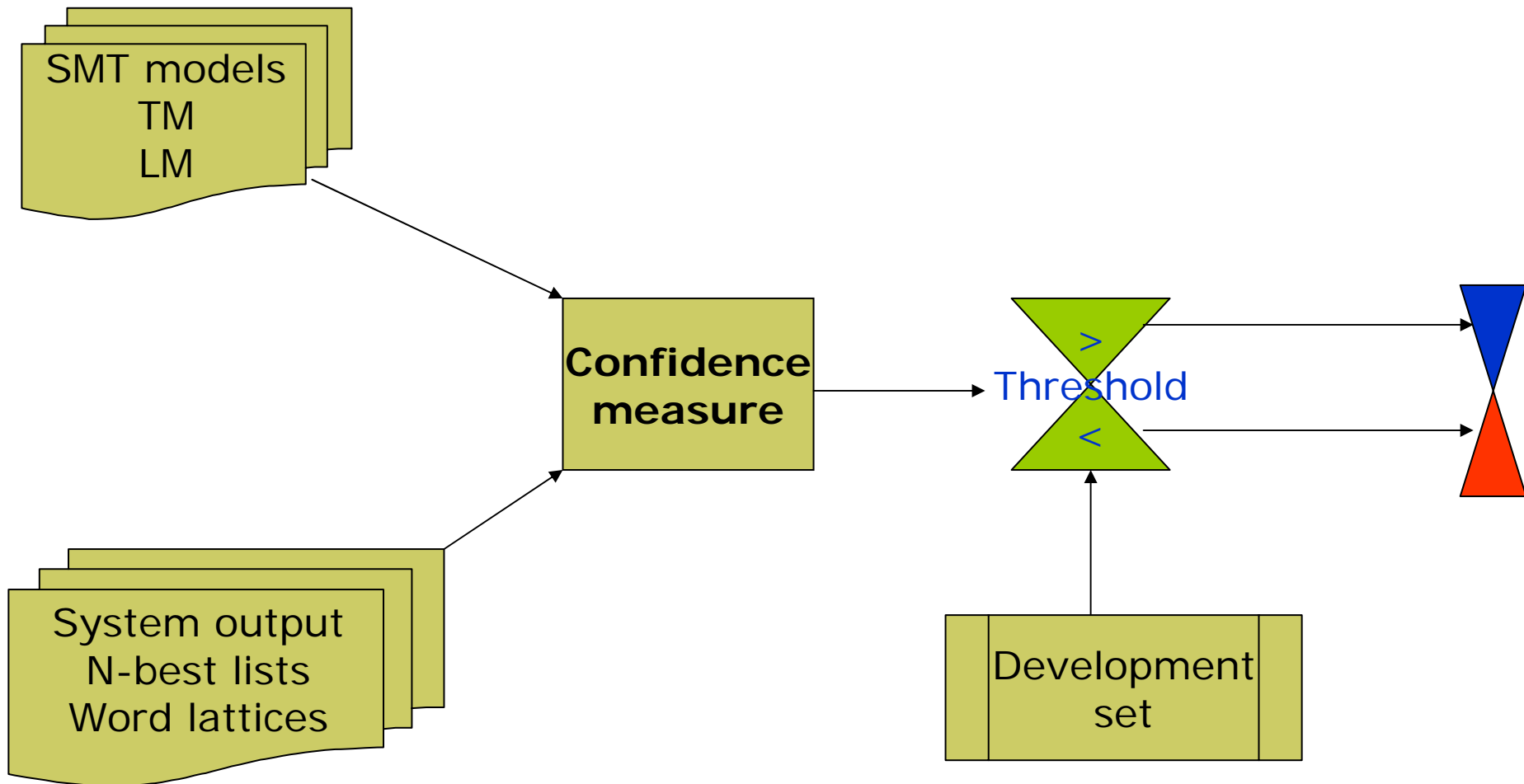
Word confidence estimation



Word confidence estimation



Word confidence estimation



An example: word posterior probability as confidence measure

FIFA/-1.513 will/-1.73212 severely/-2.21319 punish/-1.78467
fraud/-2.04383 on/-3.13695 the/-2.41013 pitch/-2.76636

An example: word posterior probability as confidence measure

FIFA/-1.513 will/-1.73212 severely/-2.21319 punish/-1.78467
fraud/-2.04383 on/-3.13695 the/-2.41013 pitch/-2.76636

Threshold: -2.3

An example: word posterior probability as confidence measure

FIFA/-1.513 will/-1.73212 severely/-2.21319 punish/-1.78467

fraud/-2.04383 on/-3.13695 the/-2.41013 pitch/-2.76636

Threshold: -2.3

An example: word posterior probability as confidence measure

FIFA/-1.513 will/-1.73212 severely/-2.21319 punish/-1.78467

fraud/-2.04383 on/-3.13695 the/-2.41013 pitch/-2.76636

Threshold: -2.3

SYS: FIFA will severely punish fraud on the pitch

The problems

- The error detection is a hard problem given very low accuracies in previous work
- Features are not adequate
 - From inner components of SMT system
 - Or from system outputs
 - Both considered by SMT system
- Need to find some **external** information sources from outside SMT system

Our solution

- Linguistic knowledge: a good external information source
 - Lexical
 - Syntactic
- Widely and successfully used in ASR error detection
 - Not widely used in SMT error detection
- Contribution: first time to successfully use linguistic features in error detection for SMT

Outline

- Linguistic features for error detection
- MaxEnt-based error detection model
- Evaluation metric
- Experiments

Outline

- Linguistic features for error detection
- MaxEnt-based error detection model
- Evaluation metrics
- Experiments

Lexical features

- **wd:** word itself
- **pos:** pos tag of words
- **Motivation:** words in frequently occurring word/pos patterns are more likely to be correct than those in rare patterns

Lexical features

- **wd:** word itself
- **pos:** pos tag of words
- **Motivation:** words in frequently occurring word/pos patterns are more likely to be correct than those in rare patterns

Although the dog

Although the off

Syntactic features

- **Motivation:** words in an ungrammatical part of a sentence are prone to be incorrect

Syntactic features

- **Motivation:** words in an ungrammatical part of a sentence are prone to be incorrect
- **Challenge:** machine generated hypotheses are rarely fully grammatical
 - Traditional constituent/dependency parsers trained on grammatical sentences are hard to deal with partially grammatical sentences

Syntactic features

- Link grammar parser
 - Produce a set of labeled links

Syntactic features

- Link grammar parser
 - Produce a set of labeled links
 - Non-link scheme
 - When the parser fails to parse the entire sentence, it ignores one word each time until it finds linkages for remaining words

Syntactic features

- Link grammar parser
 - Produce a set of labeled links
 - Non-link scheme
 - When the parser fails to parse the entire sentence, it ignores one word each time until it finds linkages for remaining words
- Ignored words are **null-linked words (disconnected islands)**: prone to be incorrect

Syntactic features

- Link grammar parser
 - Produce a set of labeled links
 - Non-link scheme
 - When the parser fails to parse the entire sentence, it ignores one word each time until it finds linkages for remaining words
- Ignored words are **null-linked words (disconnected islands)**: prone to be incorrect

$$link(w) = \begin{cases} \text{yes,} & w \text{ has links} \\ \text{no,} & \text{otherwise} \end{cases}$$

Syntactic features: examples

SRC: 沙特 为 支援 巴勒斯坦 进行 募捐

REF: Saudi Arabia conducted a fund-raising campaign for Palestinian

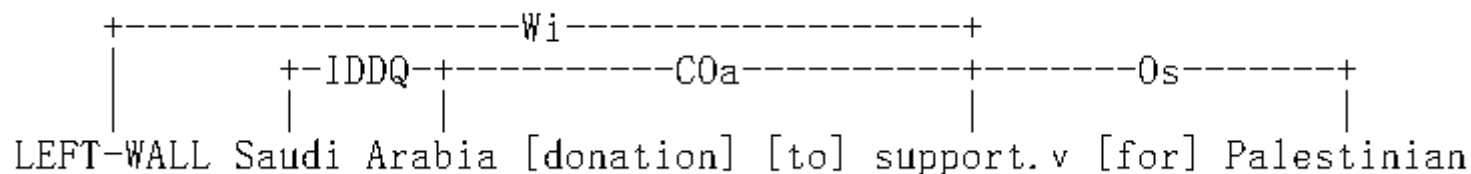
SYS: Saudi Arabia donation to support for Palestinian

Syntactic features: examples

SRC: 沙特 为 支援 巴勒斯坦 进行 募捐

REF: Saudi Arabia conducted a fund-raising campaign for Palestinian

SYS: Saudi Arabia donation to support for Palestinian

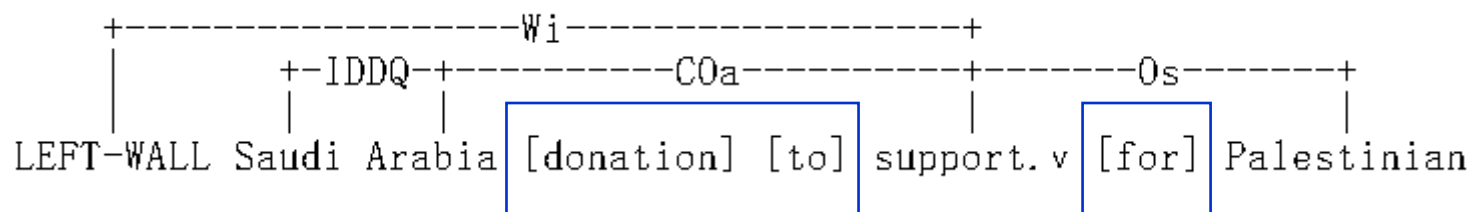


Syntactic features: examples

SRC: 沙特 为 支援 巴勒斯坦 进行 募捐

REF: Saudi Arabia conducted a fund-raising campaign for Palestinian

SYS: Saudi Arabia donation to support for Palestinian

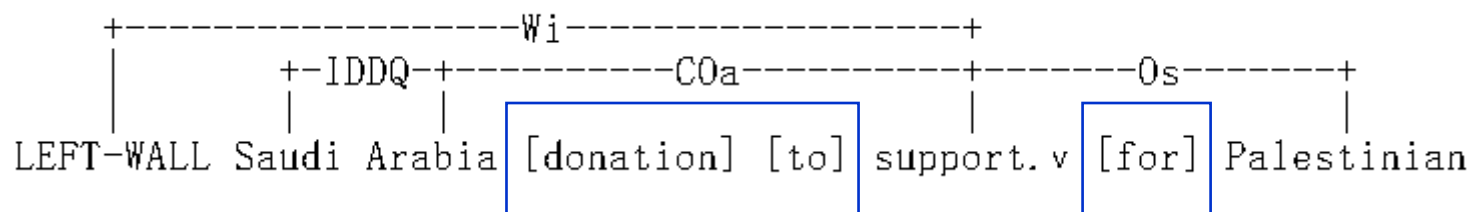


Syntactic features: examples

SRC: 沙特 为 支援 巴勒斯坦 进行 募捐

REF: Saudi Arabia conducted a fund-raising campaign for Palestinian

SYS: Saudi Arabia donation to support for Palestinian



SYS: Saudi Arabia donation to support for Palestinian

Word posterior probability

- The word posterior probability is calculated by summing up the probabilities over all hypotheses containing the word in a position which is Levenshtein aligned to the word

$$p_{wpp}(e_1^i) = \frac{\sum_{e_n: l_i(e_1, e_n) = e_1^i} p(e_n)}{\sum_1^N p(e_n)}$$

Outline

- Features for error detection
- MaxEnt-based error detection model
- Evaluation metrics
- Experiments

Error Detection Model

- MaxEnt model
- Features
 - wd
 - pos
 - link
 - dwpp (discrete word posterior probability)

Outline

- Features for error detection
- MaxEnt-based error detection model
- Evaluation metrics
- Experiments

How to determine the true class of a word in machine generated translations

- **WER**: tag a word as correct if
 - it is aligned to itself in the **Levenshtein alignment** between the hypothesis and one of reference translation
- **m-WER** (Ueffing and Ney, 2007):
 - The nearest reference translation that has minimum edit distance to the hypothesis among four reference translations.
- **PER**: tag a word as correct
 - if it occurs in one of reference translations with the same number of occurrences.

m-WER: An example

Hypothesis	China	Unicom	last	year	net	profit	rose	up	38%				
Reference	China	Unicom			net	profit	rose	up	38%	last	year		
Class	China/c	Unicom/c	last/i	year/i	net/c	profit/c	rose/c	up/c	38%/c				

m-WER: An example

Hypothesis	China	Unicom	last year	net	profit	rose	up	38%	
Reference	China	Unicom		net	profit	rose	up	38%	last year
Class	China/c	Unicom/c	last/i	year/i	net/c	profit/c	rose/c	up/c	38%/c

Evaluation metric

- CER: classification error rate
 - # Wrongly tagged words / total # words
 - Baseline CER is determined by assuming the most frequent class for all words
- Precision
 - Percentage of words correctly tagged as translation errors
- Recall
 - Percentage of actual translation errors that found by classifiers

Experiments

- SMT system: mooses
 - 100M words training corpus from LDC
 - 4-gram language model trained 180M words
 - Dev set: NIST MT-02
 - Test sets: NIST MT-03/05

Experiments

- SMT system: mooses
 - 100M words training corpus from LDC
 - 4-gram language model trained 180M words
 - Dev set: NIST MT-02
 - Test sets: NIST MT-03/05

Corpus	BLEU (%)	RCW (%)
MT-02	33.24	47.76
MT-05	32.03	47.85
MT-03	32.86	47.57

Experiments

- SMT system: mooses
 - 100M words training corpus from LDC
 - 4-gram language model trained 180M words
 - Dev set: NIST MT-02
 - Test sets: NIST MT-03/05

Corpus	BLEU (%)	RCW (%)
MT-02	33.24	47.76
MT-05	32.03	47.85
MT-03	32.86	47.57

Error detection system

	Corpus	Sentences	Words
Training	MT-02	878	24,225
Development	MT-05	1082	31,321
Test	MT-03	919	25,619

Error detection results

Combination	Features	CER (%)	Pre (%)	Rec (%)	F (%)
Baseline	-	47.57	-	-	-
Thresholding <i>wpp</i>	-	43.16	58.98	58.07	58.52
MaxEnt (<i>dwpp</i>)	44	43.07	56.12	81.86	66.59
MaxEnt (<i>wd</i>)	19,164	41.57	58.25	73.11	64.84
MaxEnt (<i>pos</i>)	199	39.90	58.88	79.23	67.55
MaxEnt (<i>link</i>)	19	44.31	54.72	89.72	67.98
MaxEnt (<i>wd + pos</i>)	19,363	39.43	59.36	78.60	67.64
MaxEnt (<i>wd + pos + link</i>)	19,382	39.79	58.74	80.97	68.08
MaxEnt (<i>dwpp + wd</i>)	19,208	41.04	57.18	83.75	67.96
MaxEnt (<i>dwpp + wd + pos</i>)	19,407	38.88	59.87	78.38	67.88
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,426	38.76	59.89	78.94	68.10

Error detection results

Combination	Features	CER (%)	Pre (%)	Rec (%)	F (%)
Baseline	-	47.57	-	-	-
Thresholding <i>wpp</i>	-	43.16	58.98	58.07	58.52
MaxEnt (<i>dwpp</i>)	44	43.07	56.12	81.86	66.59
MaxEnt (<i>wd</i>)	19,164	41.57	58.25	73.11	64.84
MaxEnt (<i>pos</i>)	199	39.90	58.88	79.23	67.55
MaxEnt (<i>link</i>)	19	44.31	54.72	89.72	67.98
MaxEnt (<i>wd + pos</i>)	19,363	39.43	59.36	78.60	67.64
MaxEnt (<i>wd + pos + link</i>)	19,382	39.79	58.74	80.97	68.08
MaxEnt (<i>dwpp + wd</i>)	19,208	41.04	57.18	83.75	67.96
MaxEnt (<i>dwpp + wd + pos</i>)	19,407	38.88	59.87	78.38	67.88
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,426	38.76	59.89	78.94	68.10

Error detection results

Combination	Features	CER (%)	Pre (%)	Rec (%)	F (%)
Baseline	-	47.57	-	-	-
Thresholding <i>wpp</i>	-	43.16	58.98	58.07	58.52
MaxEnt (<i>dwpp</i>)	44	43.07	56.12	81.86	66.59
MaxEnt (<i>wd</i>)	19,164	41.57	58.25	73.11	64.84
MaxEnt (<i>pos</i>)	199	39.90	58.88	79.23	67.55
MaxEnt (<i>link</i>)	19	44.31	54.72	89.72	67.98
MaxEnt (<i>wd + pos</i>)	19,363	39.43	59.36	78.60	67.64
MaxEnt (<i>wd + pos + link</i>)	19,382	39.79	58.74	80.97	68.08
MaxEnt (<i>dwpp + wd</i>)	19,208	41.04	57.18	83.75	67.96
MaxEnt (<i>dwpp + wd + pos</i>)	19,407	38.88	59.87	78.38	67.88
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,426	38.76	59.89	78.94	68.10

Error detection results

Combination	Features	CER (%)	Pre (%)	Rec (%)	F (%)
Baseline	-	47.57	-	-	-
Threshold	-	43.16	58.98	58.07	58.52
MaxEnt (<i>dw</i>)	44	43.07	56.12	81.86	66.59
MaxEnt (<i>wd</i>)	19,164	41.57	58.25	73.11	64.84
MaxEnt (<i>pos</i>)	199	39.90	58.88	79.23	67.55
MaxEnt (<i>link</i>)	19	44.31	54.72	89.72	67.98
MaxEnt (<i>wd + pos</i>)	19,363	39.43	59.36	78.60	67.64
MaxEnt (<i>wd + pos + link</i>)	19,382	39.79	58.74	80.97	68.08
MaxEnt (<i>dwpp + wd</i>)	19,208	41.04	57.18	83.75	67.96
MaxEnt (<i>dwpp + wd + pos</i>)	19,407	38.88	59.87	78.38	67.88
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,426	38.76	59.89	78.94	68.10

7% absolute improvement over the baseline CER

Error detection results

Combination	Features	CER (%)	Pre (%)	Rec (%)	F (%)
Baseline	-	47.57	-	-	-
Thresholding <i>wpp</i>	-	43.16	58.98	58.07	58.52
MaxEnt (<i>dwpp</i>)	44	43.07	56.12	81.86	66.59
MaxEnt (<i>wd</i>)	19,164	41.57	58.25	73.11	64.84
MaxEnt (<i>pos</i>)	199	39.90	58.88	79.23	67.55
MaxEnt (<i>link</i>)	19	44.31	54.72	89.72	67.98
MaxEnt (<i>wd + pos</i>)	19,363	39.43	59.36	78.60	67.64
MaxEnt (<i>wd + pos + link</i>)	19,382	39.79	58.74	80.97	68.08
MaxEnt (<i>dwpp + wd</i>)	19,208	41.04	57.18	83.75	67.96
MaxEnt (<i>dwpp + wd + pos</i>)	19,407	38.88	59.87	78.38	67.88
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,426	38.76	59.89	78.94	68.10

Error detection results

Combination	Features	CER (%)	Pre (%)	Rec (%)	F (%)
Baseline	-	47.57	-	-	-
Thresholding <i>wpp</i>	-	43.16	58.98	58.07	58.52
MaxEnt (<i>dwpp</i>)	44	43.07	56.12	81.86	66.59
MaxEnt (<i>wd</i>)	19,164	41.57	58.25	73.11	64.84
MaxEnt (<i>pos</i>)	199	39.90	58.88	79.23	67.55
MaxEnt (<i>link</i>)	19	44.31	54.72	89.72	67.98
MaxEnt (<i>wd + pos</i>)	19,363	39.43	59.36	78.60	67.64
MaxEnt (<i>wd + pos + link</i>)	19,382	39.79	58.74	80.97	68.08
MaxEnt (<i>dwpp + wd</i>)	19,208	41.04	57.18	83.75	67.96
MaxEnt (<i>dwpp + wd + pos</i>)	19,407	38.88	59.87	78.38	67.88
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,426	38.76	59.89	78.94	68.10

Error detection results

Combination	Features	CER (%)	Pre (%)	Rec (%)	F (%)
Baseline	-	47.57	-	-	-
Thresholding <i>wpp</i>	-	43.16	58.98	58.07	58.52
MaxEnt (<i>dwpp</i>)	44	43.07	56.12	81.86	66.59
MaxEnt (<i>wd</i>)	19,164	41.57	58.25	73.11	64.84
MaxEnt (<i>pos</i>)	199	39.90	58.88	79.23	67.55
MaxEnt (<i>link</i>)	19	44.31	54.72	89.72	67.98
MaxEnt (<i>dwpp + wd + pos</i>)	19,363	39.43	59.36	78.60	67.64
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,382	39.79	58.74	80.97	68.08
MaxEnt (<i>dwpp + wd + pos + link + pos + link</i>)	19,208	41.04	57.18	83.75	67.96
MaxEnt (<i>dwpp + wd + pos + link + pos + link + pos + link</i>)	19,407	38.88	59.87	78.38	67.88
MaxEnt (<i>dwpp + wd + pos + link</i>)	19,426	38.76	59.89	78.94	68.10

18% relative improvement over the baseline CER

Conclusions

- We have proposed a MaxEnt-based approach to automatically **detect errors** in SMT outputs
- We have integrated **linguistic features** into the error detection model

Conclusions

- We have found that
 - Linguistic features alone achieve a **higher** improvement than word posterior probabilities
 - The performance is **further improved** when we combine linguistic features with word posterior probability features.

Thank you

Questions?