Confidence measures for machine translation

Sylvain Raybaud, Caroline Lavecchia, David Langlois, Kamel Smaïli

LORIA (Nancy, France)/PAROLE group

ICAART 09
21st January 2009

A quick overview of SMT

source sentence:	$\mathbf{s} = s_1, .., s_I$
candidate translation:	$\mathbf{t} = t_1, .., t_J$

$$\hat{\mathbf{t}} = \arg\max_{\mathbf{t}} P(\mathbf{t}|\mathbf{s}) = \arg\max_{\mathbf{t}} P(\mathbf{s}|\mathbf{t}) \times P(\mathbf{t})$$

$P(\mathbf{s}|\mathbf{t})$: translation model:

$\longrightarrow$ is the idea of the source sentence kept?

$P(\mathbf{t})$: language model:

$\longrightarrow$ is the generated sentence correct?

Motivation

- Estimate **translation quality** without human references:
  - → *Costly.*
  - → *Generally not available.*
- Estimate **reliability** of each item in the translation (rather than the overall quality).
  - → *Propose a better hypothesis.*
  - → *Ask for user's verification.*

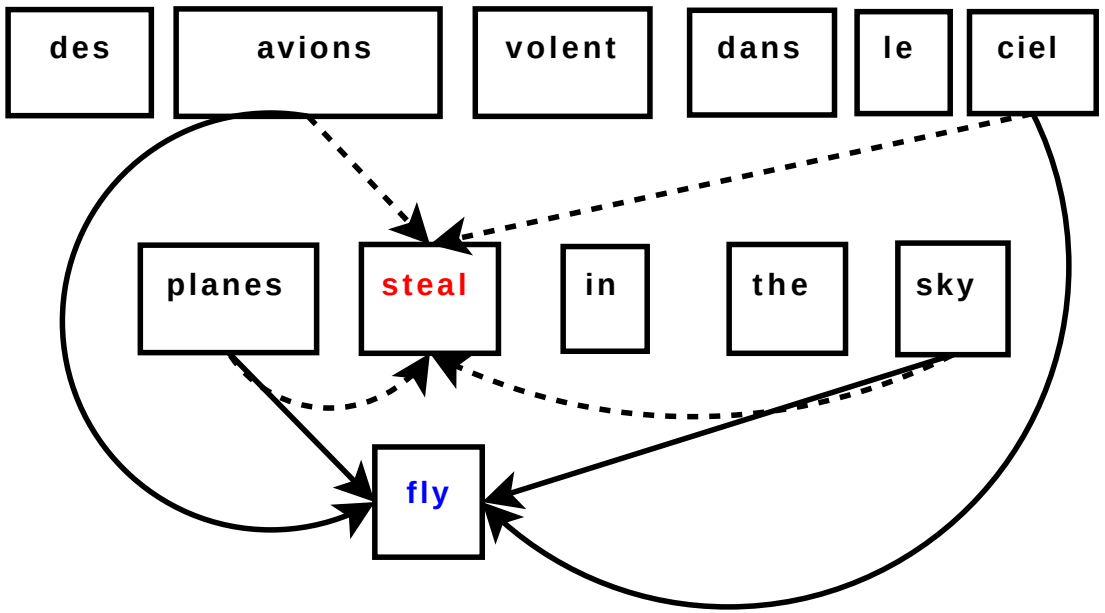
Main approaches

- *A posteriori* relevance of a word:
  - → estimate **word's probability of correctness** using the word-lattice or a translation table.
  - → compute a **score** (predictive parameter), for example likelihood ratios.
  - → **combine different measures** (perceptron, multi-layers perceptron, ...).

Our approach

Detect wrong or misplaced words by the mean of predictive parameters using the context.

Example:



Proposed measures

- We compute different measures relying on:
  - → **Mutual information** between words.
  - → **N-gram** language model.
  - → Words' **linguistic features**.
- Linear combination of the above measures.

Performance estimation

- For every target word t:

$$class(t) = \begin{cases} \text{correct} & \text{if} \quad C(t) > \delta \\ \text{incorrect} & \text{if} \quad C(t) < \delta \end{cases}$$

- Machine classification compared to man-made one.
- Performance = reliability of the classification.

Example:

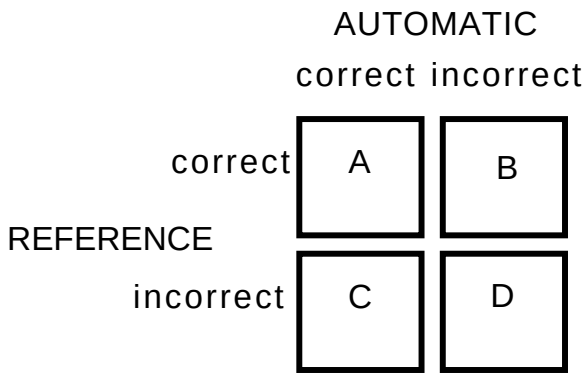
source	des	avions	volent	dans	le	ciel
translation	a	planes	steal	in	the	sky
human	0	1	0	1	1	1
machine	1	0	0	1	1	1

correct acceptances: 3

correct rejections: 1

Evaluation metrics

- **ROC curve**, **F-measure** and **Error Rate**:



correct acceptances: **A**

correct rejections: **D**

incorrect acceptances: C

incorrect rejections: B

Error Rate (ER): $\frac{B+C}{A+B+C+D}$

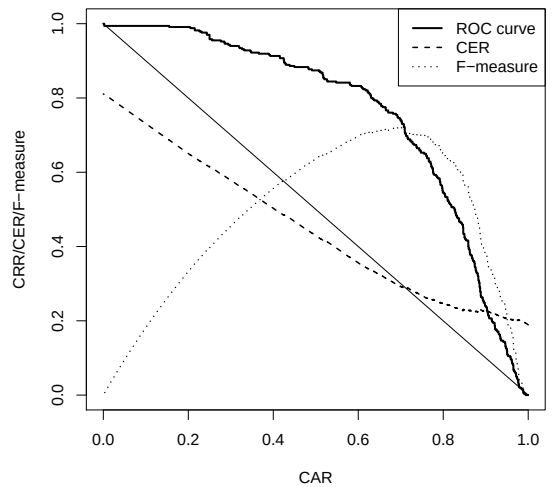
Correct Acceptance Rate (CAR): $\frac{A}{A+B}$ — proportion of correct words which are actually accepted.

Correct Rejection Rate (CRR): $\frac{D}{C+D}$ — proportion of incorrect words which are actually rejected.

F-measure: $\frac{2 \times CAR \times CRR}{CAR+CRR}$

Receiver Operating Characteristics curve

- Given threshold $\longrightarrow$ determines (CAR, CRR) $\longrightarrow$ a point.
- Testing different thresholds $\longrightarrow$ **ROC curve**.
- As the threshold increases, CRR goes from 0 to 1 and CAR from 1 to 0.



- Very easy interpretation: **the closer the curve is to** $(1, 1)$**, the better.**

Experimental settings

English to French phrase based translation	
Language model	SRILM toolkit, 1 to 5-grams models
Translation model	GIZA++, IBM-5 model
Decoder	Pharaoh
Corpora	EUROPARL 2005 (500k English-French sentences pairs)

- Confidence measures evaluated on the test corpus (around 5,000 words).

Mutual information

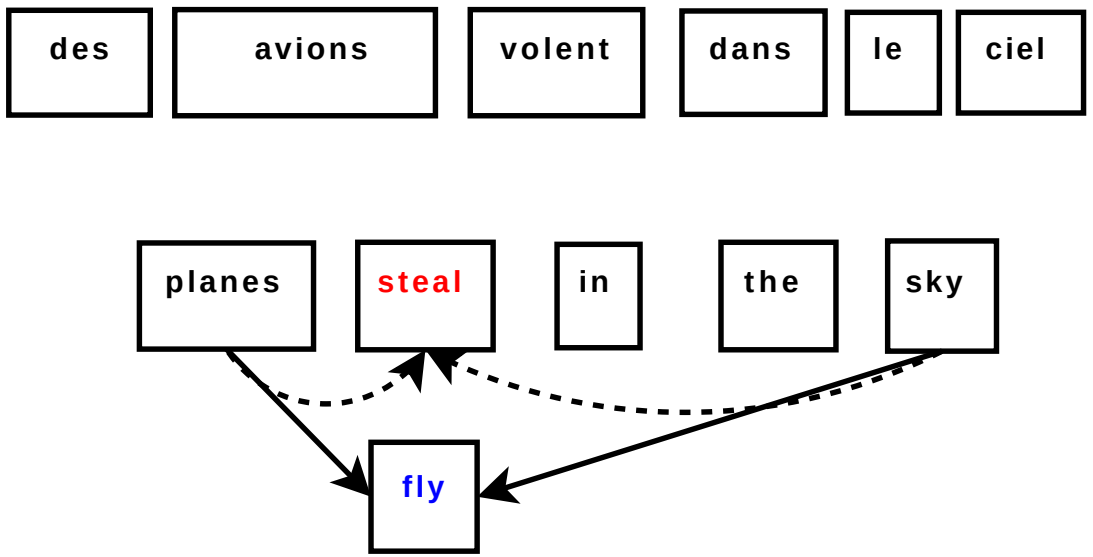
- Measures the *dependency* between two random variables.

$$\mathcal{I}(X,Y) = \sum_x \sum_y P(X=x, Y=y) log_2 \left(\frac{P(X=x, Y=y)}{P(X=x)P(Y=y)} \right)$$

- In statistical linguistic: detects words that co-occur.

Intra-language MI

- Evaluates the relevance of a word given its context.



How do we do that?

Training

$$
\begin{aligned}
p(x) &= \text{probability that word } \mathbf{x} \text{ occurs in a sentence.} \\
p(x,y) &= \text{probability that words } \mathbf{x} \text{ and } \mathbf{y} \\
&\quad \text{appear in the same sentence.}
\end{aligned}
$$

We consider the contribution of each word pair separately:

$$\mathcal{I}(x,y) = p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Training (continued): Intra-language triggers

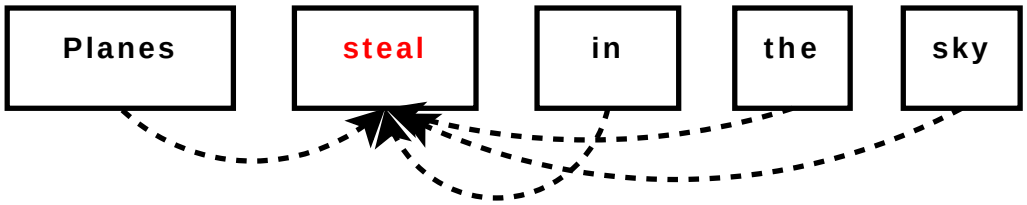
We get a so called **triggers list**:

word X triggers word Y	relationship is quantified by *mutual information*

Example:

security	$\leftrightarrow$	*foreign*	$6.58 \cdot 10^{-3}$
security	$\leftrightarrow$	*policy*	$6.17 \cdot 10^{-3}$
security	$\leftrightarrow$	*social*	$4.33 \cdot 10^{-3}$
	...		
policy	$\leftrightarrow$	*common*	$1.12 \cdot 10^{-2}$
policy	$\leftrightarrow$	*foreign*	$1.05 \cdot 10^{-2}$
policy	$\leftrightarrow$	*agricultural*	$6.90 \cdot 10^{-3}$

Word's confidence estimation



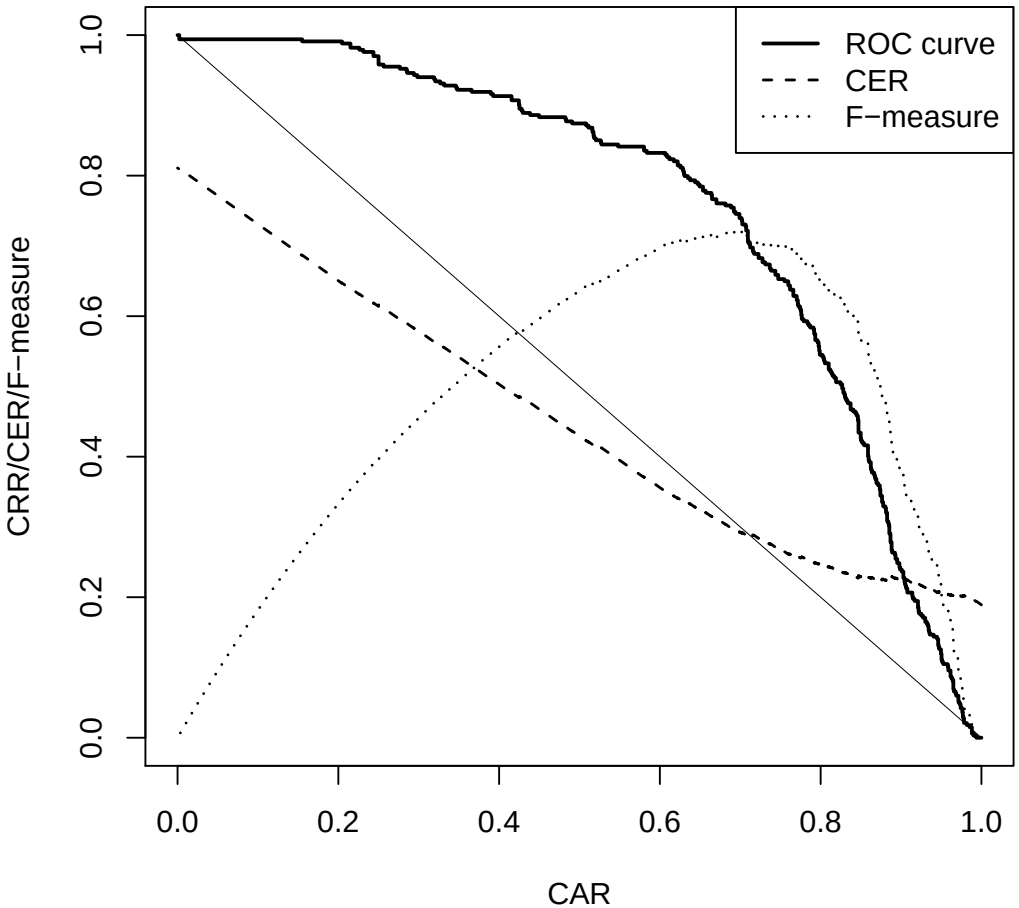
- Word's score: average mutual information between this word and the others in the sentence:

$$\begin{aligned} \mathbf{t} &= t_1, .., t_J \\ C(t_j) &= \frac{1}{J-1} \sum_{i=1,..,J \ i \neq j} \mathcal{I}(t_i, t_j) \end{aligned}$$

Possible tweaks:

- Take words' positions into account.
- Ignore function words (*the, of, ...*).

Efficiency



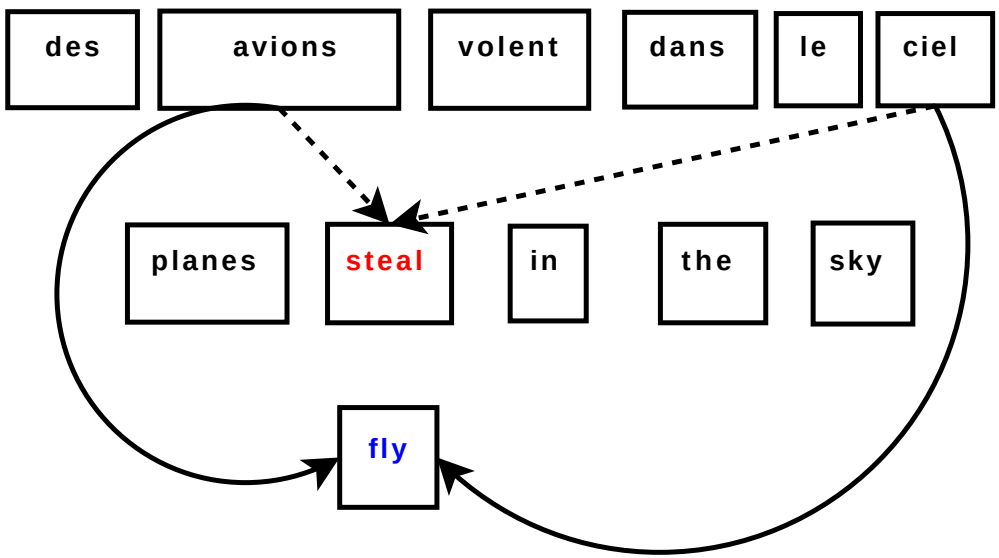
- Words' positions not taken into account.
- function words ignored.

Best classifier wrt. F-measure:

- F-measure: 0.72
- Error Rate: 0.30
- CAR: 0.69
- CRR: 0.75

Inter-language MI

- Detects inconsistent translations.



How do we do that?

Training

We are looking for couples (x, y) such that $x \in$ source and $y \in$ translation

$$
\begin{aligned}
p_S(x) &= \text{probability that } \mathbf{x} \text{ occurs in a } \textbf{source} \text{ sentence} \\
p_T(y) &= \text{probability that } \mathbf{y} \text{ occurs in a } \textbf{translated} \text{ sentence} \\
p(x,y) &= \text{probability that } \mathbf{x} \text{ occurs in a } \textbf{source} \\
&\quad \text{and } \mathbf{y} \text{ in its } \textbf{translation}
\end{aligned}
$$

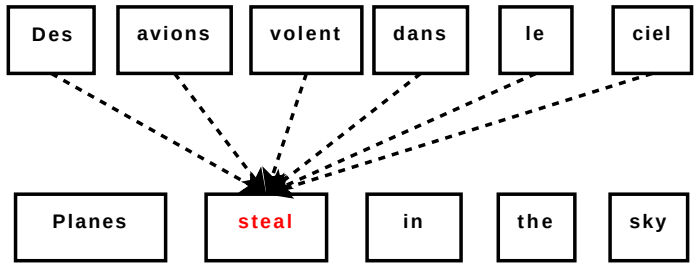
$$
\mathcal{I}(x,y) = p(x,y) \log \left(\frac{p(x,y)}{p_S(x) p_T(y)} \right)
$$

Remark: Can be used to build translation tables.

Training (continued): Inter-languages triggers

sécurité	$\longrightarrow$	*security*	$4.39 \cdot 10^{-2}$
sécurité	$\longrightarrow$	*safety*	$3.76 \cdot 10^{-2}$
sécurité	$\longrightarrow$	*foreign*	$5.65 \cdot 10^{-3}$
	...		
politique	$\longrightarrow$	*policy*	$1.39 \cdot 10^{-1}$
politique	$\longrightarrow$	*political*	$5.75 \cdot 10^{-2}$
politique	$\longrightarrow$	*common*	$1.23 \cdot 10^{-2}$

Word's confidence estimation



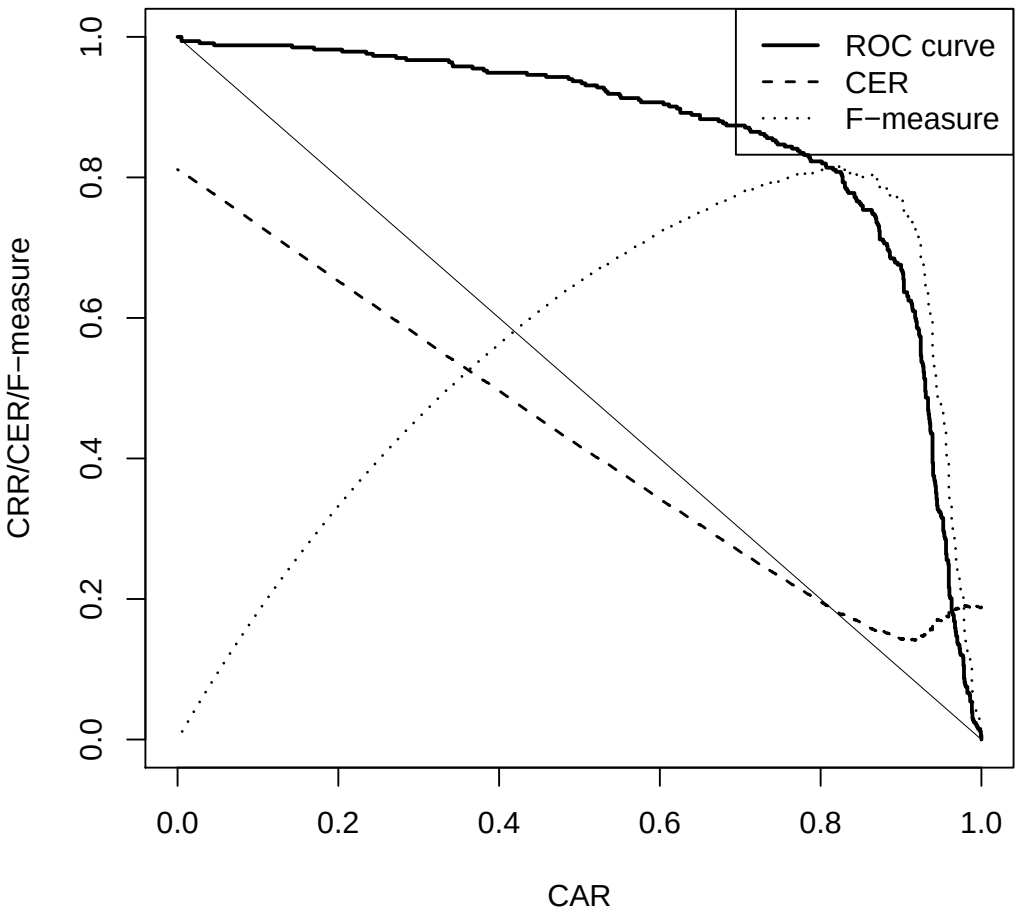
- A word's score is the average MI between this word and the words in the source sentence:

$$\begin{aligned} \mathbf{s} &= s_1, .., s_I \\ \mathbf{t} &= t_1, .., t_J \\ C(t_j) &= \frac{1}{I} \sum_{i=1,..,I} \mathcal{I}(s_i, t_j) \end{aligned}$$

Same possible tweaks:

- Consider words positions.
- Ignore function words.

Efficiency



- Limited distorsion.
- function words ignored.

Best classifier wrt. F-measure:

- F-measure: 0.82
- Error Rate: 0.18
- CAR: 0.83
- CRR: 0.81

N-gram based measure

The decoder makes a tradoff between $P(\mathbf{t})$ and $P(\mathbf{s}|\mathbf{t})$; unbalanced choices penalised:

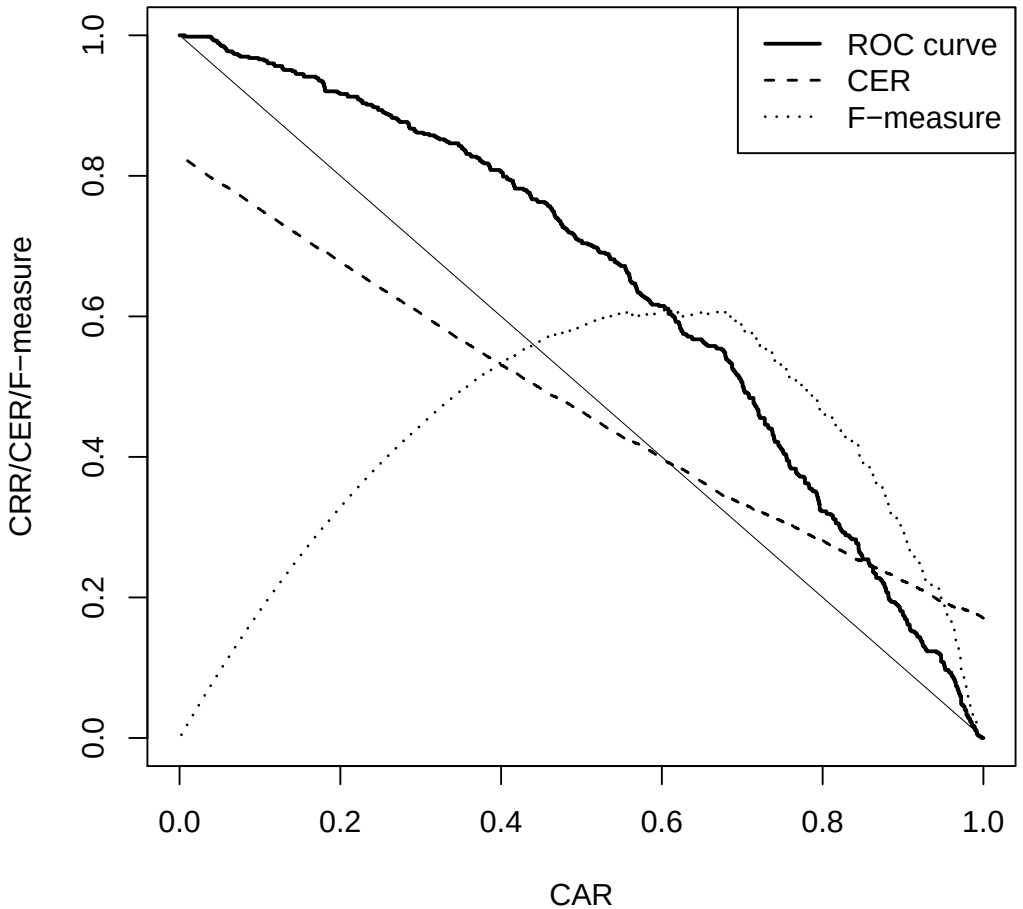
$$P(\mathbf{t}) \text{ too low} \rightarrow \mathbf{t} \text{ often incorrect}$$

$$\begin{aligned} \mathbf{t} &= t_1, .., t_J \\ C(t_j) &= P(t_j | t_{j-n+1}, .., t_{j-1}) \end{aligned}$$

Example:

$$\begin{cases} \text{P(steal} \mid \text{the planes)} = 0.001 \\ \text{P(fly} \mid \text{the planes)} = 0.4 \end{cases}$$

Efficiency



- Interesting performances but not as good as MI's.

Best classifier wrt. F-measure:

- F-measure: 0.61
- Error Rate: 0.39
- CAR: 0.61
- CRR: 0.61

Linguistic Features

- Directly detect **grammatical errors**.
- Replace each word by a set of features: grammatical category, tense, person, gender, number.
- Features given by BDLEX.

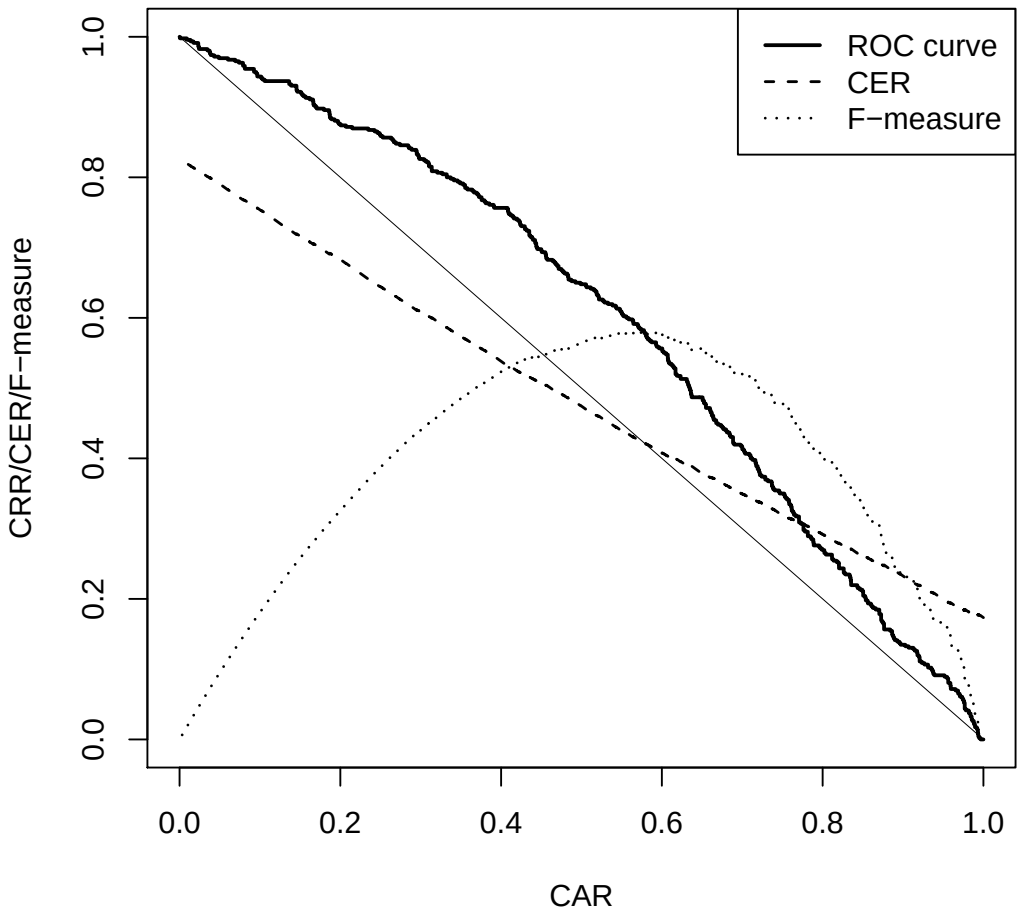
Examples:

$t =$was	$\rightarrow$	$\tilde{t} =$*verb, indicative imperfect, 3rd person*
$t =$word	$\rightarrow$	$\tilde{t} =$*noun, masculine, singular*

N-gram features model:

$$C(t_j) = P(\tilde{t}_j | \tilde{t}_{n-j+1}, .., \tilde{t}_{j-1})$$

Efficiency



- Inefficient when used alone.
- Disambiguation is difficult.

Best classifier wrt. F-measure:

- F-measure: 0.58
- Error Rate: 0.43
- CAR: 0.57
- CRR: 0.59

Motivation

- Each measure detects a specific kind of error.
- Combining them yields more powerful classifiers.

Perceptron

- Scores are **linearly combined**:

$$C(t) = \sum_{i=1}^{k} w_i C_i(t)$$

- **Weights are optimised** by comparing expected class $c_{reference}$ to produced one c:

$$\forall t \ \forall i \ w_i \leftarrow w_i + \alpha \left(c_{reference}(t) - c(t)\right) C_i(t)$$

Efficiency

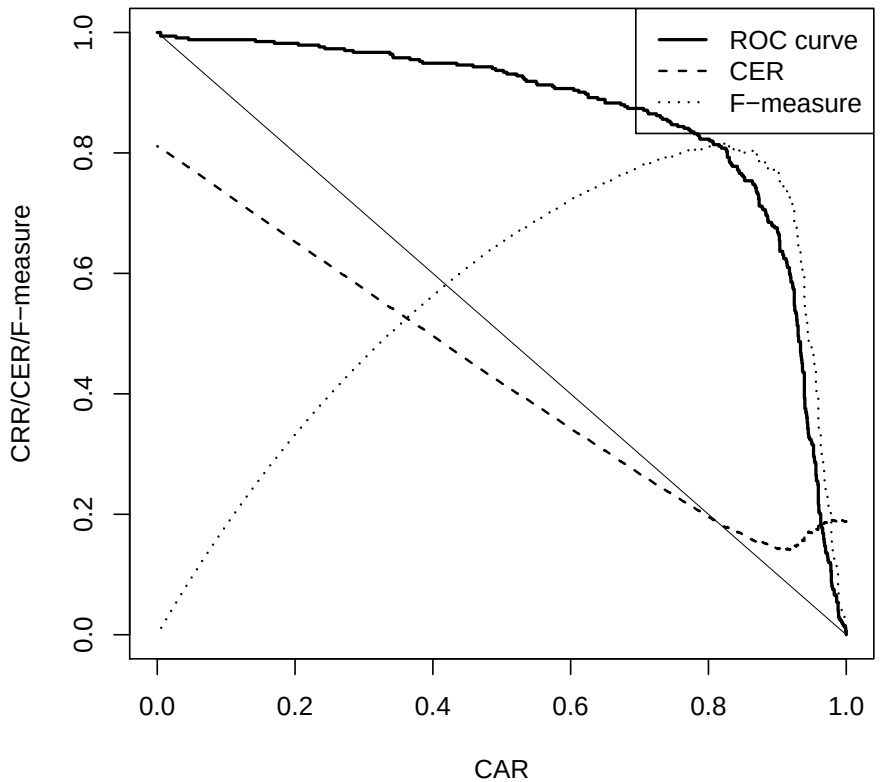


Figure: Inter-MI alone

F-measure: 0.82 - CAR: 0.83 - CRR: 0.81 - ER: 0.18

Performances

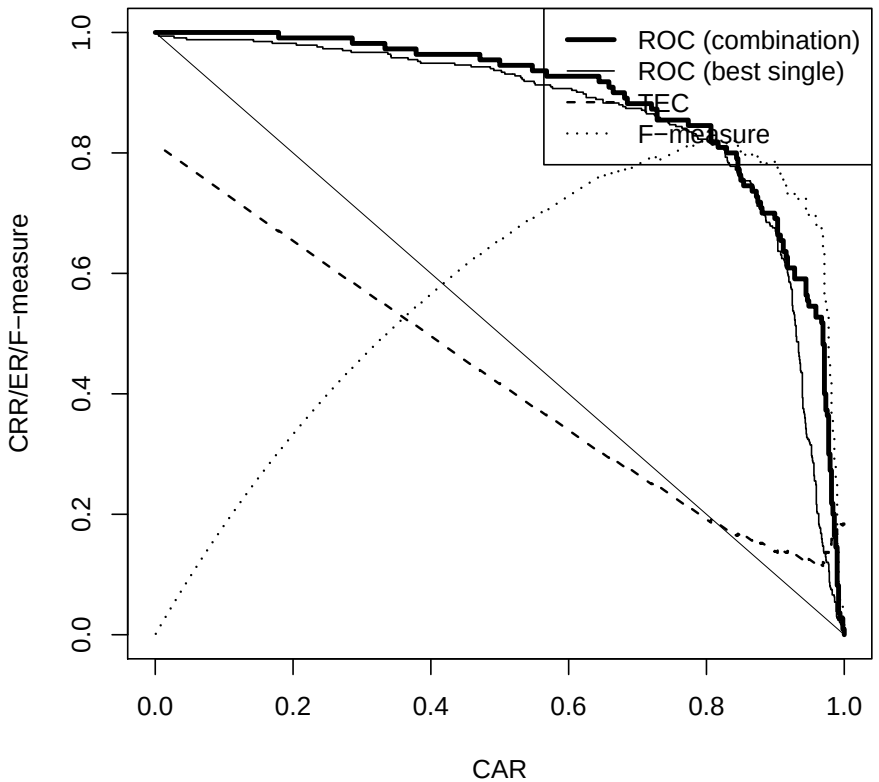


Figure: Intra- and inter-MI combined

F-measure: 0.83 - CAR: 0.81 - CRR: 0.85 - ER: 0.19

Conclusion

- We propose four original confidence measures (plus combination).
- Some achieve results comparable to state-of-the-art.

Future work

- Adapt more measures from ASR.
- Propose new original measures for MT.
- Come up with a smarter combination method.
- More thorough performance evaluation.

- Questions!