Using Language Models to Assist in Correction of Machine Translation Output

Beatrice Alex
MSc in Speech and Language Processing
University of Edinburgh
June 2002

Supervisors:
Miles Osborne
Donnla Nic Gearailt

Today’s Presentation

Today’s Presentation

- Machine translation (MT) versus Machine-assisted translation (MAT)

Today’s Presentation

- Machine translation (MT) versus Machine-assisted translation (MAT)
- Recent work in automatic MT evaluation

Today’s Presentation

- Machine translation (MT) versus Machine-assisted translation (MAT)
- Recent work in automatic MT evaluation
- My project:

Today's Presentation

- Machine translation (MT) versus Machine-assisted translation (MAT)
- Recent work in automatic MT evaluation
- My project:
  - Resources

Today's Presentation

- Machine translation (MT) versus Machine-assisted translation (MAT)
- Recent work in automatic MT evaluation
- My project:
  - Resources
  - Goals and sub-goals

Today's Presentation

- Machine translation (MT) versus Machine-assisted translation (MAT)
- Recent work in automatic MT evaluation
- My project:
  - Resources
  - Goals and sub-goals
  - Preliminary results

Today's Presentation

- Machine translation (MT) versus Machine-assisted translation (MAT)
- Recent work in automatic MT evaluation
- My project:
  - Resources
  - Goals and sub-goals
  - Preliminary results
  - Potential difficulties

A Critical Look at MT/ 1

A Critical Look at MT/ 1

- MT is error-prone due to:

A Critical Look at MT/ 1

- MT is error-prone due to:
  - Grammatical and syntactic differences between languages

A Critical Look at MT/ 1

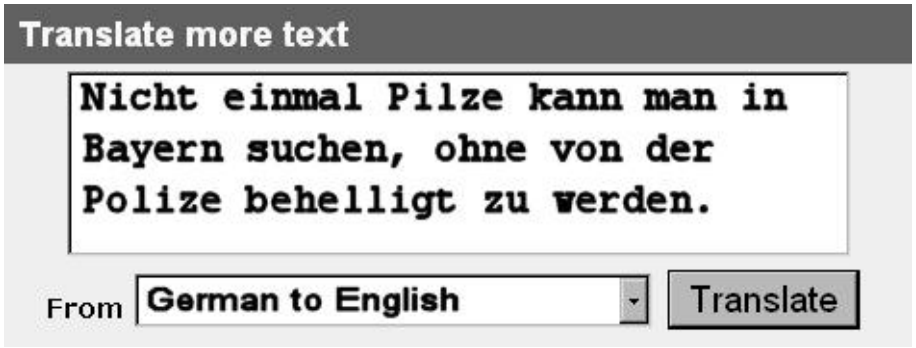
- MT is error-prone due to:
  - Grammatical and syntactic differences between languages
  - Cultural differences and world knowledge

A Critical Look at MT/ 1

- MT is error-prone due to:
  - Grammatical and syntactic differences between languages
  - Cultural differences and world knowledge
  - Collocations and idiomatic expressions, e.g. “kick the bucket”

A Critical Look at MT/ 1

- MT is error-prone due to:
  - Grammatical and syntactic differences between languages
  - Cultural differences and world knowledge
  - Collocations and idiomatic expressions, e.g. “kick the bucket”



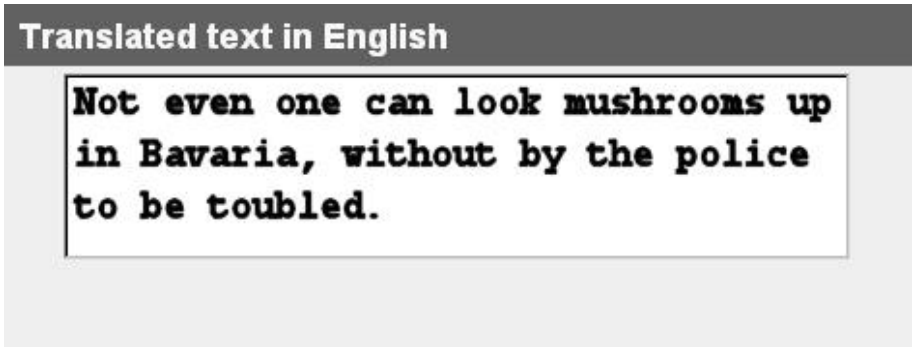


Figure 1 and 2: German utterance translated by Systran into English

A Critical Look at MT/ 2

A Critical Look at MT/ 2

- Spotting and correcting errors by hand is time- and labour-consuming

A Critical Look at MT/ 2

- Spotting and correcting errors by hand is time- and labour-consuming
- Is an automatic approach more desirable than improving the MT engine itself?

Machine-Assisted Translation versus MT

Machine-Assisted Translation versus MT

- Refining MT engines is:

Machine-Assisted Translation versus MT

- Refining MT engines is:
  - source language (SL) and target language (TL) dependent

Machine-Assisted Translation versus MT

- Refining MT engines is:
  - source language (SL) and target language (TL) dependent
  - complicated and time-consuming

Machine-Assisted Translation versus MT

- Refining MT engines is:
  - source language (SL) and target language (TL) dependent
  - complicated and time-consuming
- A **fast** and **cheap** error spotting in MT output - NOW

Machine-Assisted Translation versus MT

- Refining MT engines is:
  - source language (SL) and target language (TL) dependent
  - complicated and time-consuming
- A **fast** and **cheap** error spotting in MT output - NOW

  $\rightarrow$ attractive for system designers and end users.

Machine-Assisted Translation versus MT

- Refining MT engines is:
  - source language (SL) and target language (TL) dependent
  - complicated and time-consuming
- A **fast** and **cheap** error spotting in MT output - NOW

  $\rightarrow$ attractive for system designers and end users.
- How can MT output be automatically critiqued?

Automatic Selection of the Best Output from Multiple MT Engines/ 1

(Callison-Burch, C., Flournoy, R.S., 2001)

Automatic Selection of the Best Output from Multiple MT Engines/ 1

(Callison-Burch, C., Flournoy, R.S., 2001)

- Assumption: most fluent is best

Automatic Selection of the Best Output from Multiple MT Engines/ 1

(Callison-Burch, C., Flournoy, R.S., 2001)

- Assumption: most fluent is best
- Statistical language model (SLM) to rank translations

Automatic Selection of the Best Output from Multiple MT Engines/ 1

(Callison-Burch, C., Flournoy, R.S., 2001)

- Assumption: most fluent is best
- Statistical language model (SLM) to rank translations
- Evaluation against human-ranked data

Automatic Selection of the Best Output from Multiple MT Engines/ 1

(Callison-Burch, C., Flournoy, R.S., 2001)

- Assumption: most fluent is best
- Statistical language model (SLM) to rank translations
- Evaluation against human-ranked data
- Results: improvement of 19% compared to baseline

Automatic Selection of the Best Output from Multiple MT Engines/ 2

Automatic Selection of the Best Output from Multiple MT Engines/ 2

- Advantages:

Automatic Selection of the Best Output from Multiple MT Engines/ 2

- Advantages:
  - SLM represents probability distribution over sequences of words in the corpus

Automatic Selection of the Best Output from Multiple MT Engines/ 2

- Advantages:
  - SLM represents probability distribution over sequences of words in the corpus
  - SL independent approach

Automatic Selection of the Best Output from Multiple MT Engines/ 2

- Advantages:
  - SLM represents probability distribution over sequences of words in the corpus
  - SL independent approach
  - Simple to compute and employ

BLEU: Automatic MT Evaluation via SLMs/ 1

(Papineni et al., 2001)

BLEU: Automatic MT Evaluation via SLMs/ 1

(Papineni et al., 2001)

- Assumption: the closer MT is to human translations the better

BLEU: Automatic MT Evaluation via SLMs/ 1

(Papineni et al., 2001)

- Assumption: the closer MT is to human translations the better
- Test corpus: translations produced by different MT engines

BLEU: Automatic MT Evaluation via SLMs/ 1

(Papineni et al., 2001)

- Assumption: the closer MT is to human translations the better
- Test corpus: translations produced by different MT engines
- Reference: a corpus of expert translations

BLEU: Automatic MT Evaluation via SLMs/ 1

(Papineni et al., 2001)

- Assumption: the closer MT is to human translations the better
- Test corpus: translations produced by different MT engines
- Reference: a corpus of expert translations
- Translation quality: distance between each test candidate and its ideal reference translations

BLEU: Automatic MT Evaluation via SLMs/ 1

(Papineni et al., 2001)

- Assumption: the closer MT is to human translations the better
- Test corpus: translations produced by different MT engines
- Reference: a corpus of expert translations
- Translation quality: distance between each test candidate and its ideal reference translations
- Sentence brevity penalty

BLEU: Automatic MT Evaluation via SLMs/ 1

(Papineni et al., 2001)

- Assumption: the closer MT is to human translations the better
- Test corpus: translations produced by different MT engines
- Reference: a corpus of expert translations
- Translation quality: distance between each test candidate and its ideal reference translations
- Sentence brevity penalty
- Human evaluation: by monolingual and bilingual group

BLEU: Automatic MT Evaluation via SLMs/ 2

BLEU: Automatic MT Evaluation via SLMs/ 2

- Results: ability to estimate small differences in translation quality between MT engines

BLEU: Automatic MT Evaluation via SLMs/ 2

- Results: ability to estimate small differences in translation quality between MT engines

- Advantages: quick, inexpensive and SL independent approach

BLEU: Automatic MT Evaluation via SLMs/ 2

- Results: ability to estimate small differences in translation quality between MT engines
- Advantages: quick, inexpensive and SL independent approach
- Cost: set of professional human translations needed to train the LM

Machine Translation At Present

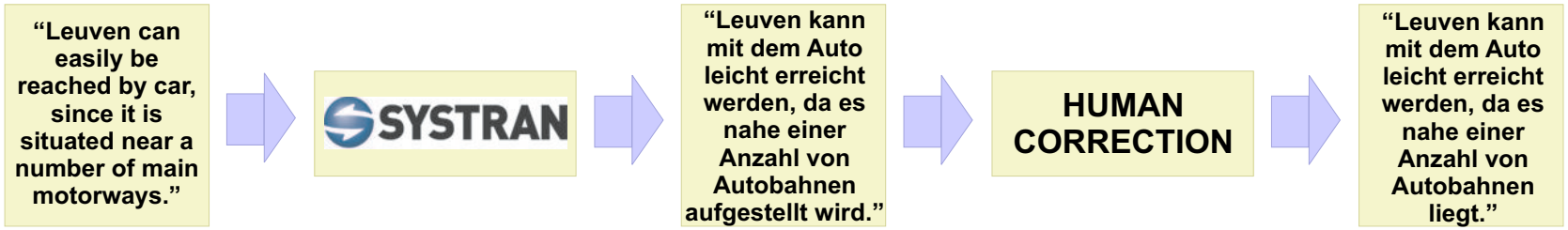


Figure 3: MT output containing errors

Machine Translation At Present

Figure 4: MT output with highlighted errors

Machine-Assisted Translation

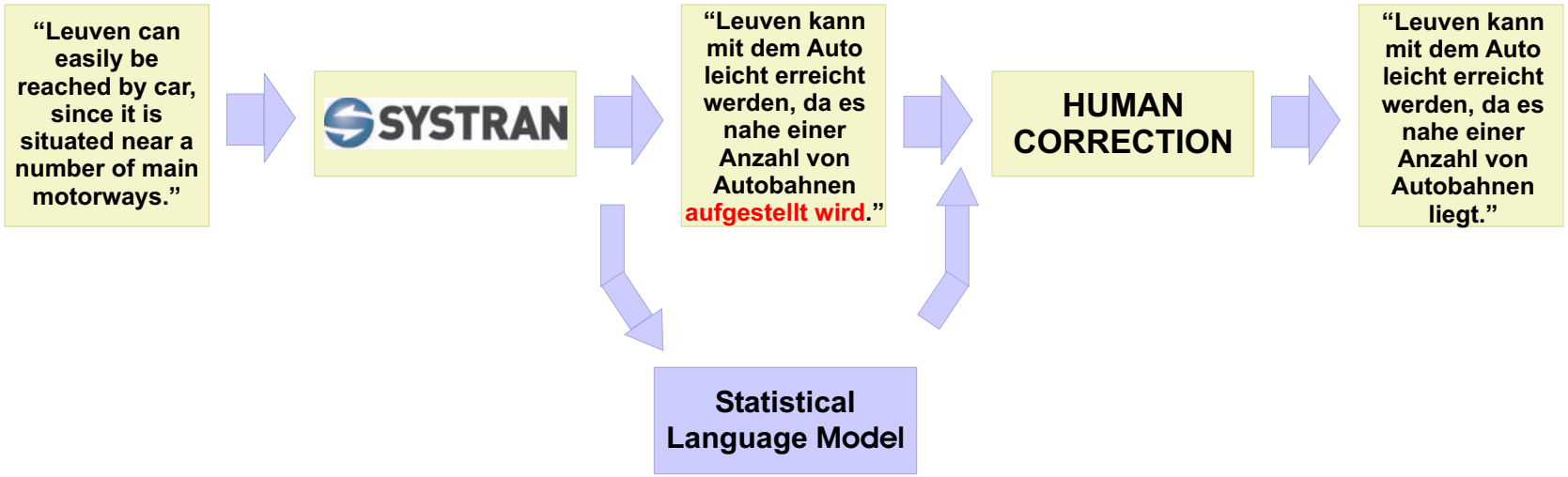


Figure 5: MAT by means of SLMs

My approach

My approach

- **Main objectives:**

My approach

- **Main objectives:**

  $\rightarrow$ *Automatically spot sentences containing translation errors in MT output by means of SLMs built from a German corpus*

My approach

- **Main objectives:**

  $\rightarrow$ *Automatically spot sentences containing translation errors in MT output by means of SLMs built from a German corpus*

  $\rightarrow$ *Differentiate between good and bad quality translated sentences in test sets while taking sentence length into account*

My approach

- **Main objectives:**

  $\rightarrow$ *Automatically spot sentences containing translation errors in MT output by means of SLMs built from a German corpus*

  $\rightarrow$ *Differentiate between good and bad quality translated sentences in test sets while taking sentence length into account*

- **Ultimate goal:**

My approach

- **Main objectives:**

  $\rightarrow$ *Automatically spot sentences containing translation errors in MT output by means of SLMs built from a German corpus*

  $\rightarrow$ *Differentiate between good and bad quality translated sentences in test sets while taking sentence length into account*

- **Ultimate goal:**

  $\rightarrow$ *Determine threshold and spot errors*

Anticipated Results

Anticipated Results

- High-scoring test translation matches word sequence in training corpus in word choice and word order

Anticipated Results

- High-scoring test translation matches word sequence in training corpus in word choice and word order

- Hypotheses:

Anticipated Results

- High-scoring test translation matches word sequence in training corpus in word choice and word order

- Hypotheses:
  - Highest scores: newspaper text

Anticipated Results

- High-scoring test translation matches word sequence in training corpus in word choice and word order

- Hypotheses:
  - Highest scores: newspaper text
  - Lowest scores: non-newspaper text

Anticipated Results

- High-scoring test translation matches word sequence in training corpus in word choice and word order

- Hypotheses:
  - Highest scores: newspaper text
  - Lowest scores: non-newspaper text
  - But quality ranking of translated sentences within sets is expected to be similarly divided into good and bad

Resources

Resources

- Machine translation engine: Systran

Resources

- Machine translation engine: Systran
- Language model: CMU Statistical Language Modelling Toolkit

Resources

- Machine translation engine: Systran
- Language model: CMU Statistical Language Modelling Toolkit
- Training data: 90 million words of German newspaper text

Resources

- Machine translation engine: Systran
- Language model: CMU Statistical Language Modelling Toolkit
- Training data: 90 million words of German newspaper text
- Test sets: English test sets from different domains

Resources

- Machine translation engine: Systran
- Language model: CMU Statistical Language Modelling Toolkit
- Training data: 90 million words of German newspaper text
- Test sets: English test sets from different domains
  $\rightarrow$ English newspaper text from The Washington Post

Resources

- Machine translation engine: Systran
- Language model: CMU Statistical Language Modelling Toolkit
- Training data: 90 million words of German newspaper text
- Test sets: English test sets from different domains
  $\rightarrow$ English newspaper text from The Washington Post
  $\rightarrow$ English newspaper text from FAZ

Resources

- Machine translation engine: Systran
- Language model: CMU Statistical Language Modelling Toolkit
- Training data: 90 million words of German newspaper text
- Test sets: English test sets from different domains
  - $\rightarrow$ English newspaper text from The Washington Post
  - $\rightarrow$ English newspaper text from FAZ
  - $\rightarrow$ Text from other domains

Resources

- Machine translation engine: Systran
- Language model: CMU Statistical Language Modelling Toolkit
- Training data: 90 million words of German newspaper text
- Test sets: English test sets from different domains
  - $\rightarrow$ English newspaper text from The Washington Post
  - $\rightarrow$ English newspaper text from FAZ
  - $\rightarrow$ Text from other domains
- Good quality human translations

What needs to be done?

What needs to be done?

- Prepare data for LM: tokenising

What needs to be done?

- Prepare data for LM: tokenising
- Train LM on tokenised training data

What needs to be done?

- Prepare data for LM: tokenising
- Train LM on tokenised training data
- Rank sentences of machine translated test sets according to the log probabilities produced by the LM

What needs to be done?

- Prepare data for LM: tokenising
- Train LM on tokenised training data
- Rank sentences of machine translated test sets according to the log probabilities produced by the LM
- Use professional human translations as a ’reference point’ to the quality of the LM

What needs to be done?

- Prepare data for LM: tokenising
- Train LM on tokenised training data
- Rank sentences of machine translated test sets according to the log probabilities produced by the LM
- Use professional human translations as a ’reference point’ to the quality of the LM
- Evaluate and revise LM

CMU-SLM Toolkit

CMU-SLM Toolkit

- Estimates probability distribution over sequences of words

CMU-SLM Toolkit

- Estimates probability distribution over sequences of words
- **Markov assumption:** only prior local context affects the next word

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-1})$$

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-2}, \ldots, W_{n-1})$$

for a given word sequence $W = (w_1, \ldots, w_n)$

CMU-SLM Toolkit

- Estimates probability distribution over sequences of words
- **Markov assumption:** only prior local context affects the next word

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-1})$$

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-2}, \ldots, W_{n-1})$$

for a given word sequence $W = (w_1, \ldots, w_n)$

- What about unseen and sparse events in training and test data?

CMU-SLM Toolkit

- Estimates probability distribution over sequences of words
- **Markov assumption:** only prior local context affects the next word

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-1})$$

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-2}, \ldots, W_{n-1})$$

for a given word sequence $W = (w_1, \ldots, w_n)$

- What about unseen and sparse events in training and test data?

  $\rightarrow$ Restricting size of training vocabulary

CMU-SLM Toolkit

- Estimates probability distribution over sequences of words
- **Markov assumption:** only prior local context affects the next word

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-1})$$

$$P(W) \approx \prod_{n=1}^{N} P(W_n|W_{n-2}, \ldots, W_{n-1})$$

for a given word sequence $W = (w_1, \ldots, w_n)$

- What about unseen and sparse events in training and test data?

  $\rightarrow$ Restricting size of training vocabulary
  $\rightarrow$ **Backing-off** and **smoothing**

Preliminary Results/ 1

Preliminary Results/ 1

- Tested 40 MT sentences all of same length and corresponding human translations

Preliminary Results/ 1

- Tested 40 MT sentences all of same length and corresponding human translations

- Results: majority of human translations received lower perplexity/ entropy than machine translations

Preliminary Results/ 1

- Tested 40 MT sentences all of same length and corresponding human translations

- Results: majority of human translations received lower perplexity/ entropy than machine translations

Original: It is conceivable that it won't be any different this time.			
Machine	**Perplexity/Entropy**	**Human**	**Perplexity/Entropy**
<s> ES IST DENKBAR DASS SIE DIESESMAL KEIN UNTERSCHIEDLICHES IST </s>	**PP = 458.06 H = 8.85 bits**	**<s> ES IST DENKBAR DASS ES DIESESMAL NICHT ANDERS WIRD </S>**	**PP = 86.15 H = 6.43 bits**

Figure 6: An Example

Preliminary Results/ 2

Preliminary Results/ 2

- Tested English/ German parallel corpus:

Preliminary Results/ 2

- Tested English/ German parallel corpus:

  → Whole chunk:

  German test set → PP = 530.67, H= 9.05 bits
  MT test set → PP = 1100.70, H = 10.10 bits

Preliminary Results/ 3

Preliminary Results/ 3

→ Individual sentences:

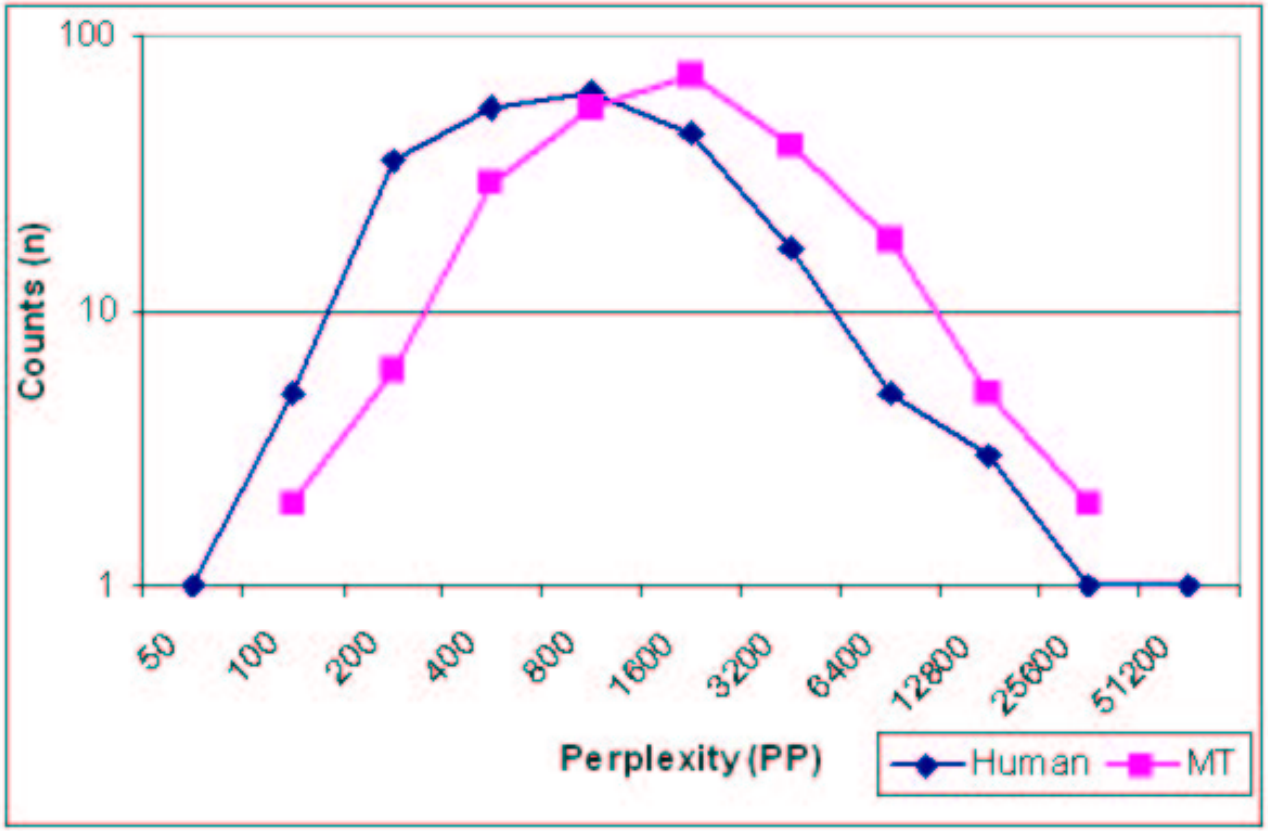


Figure 7: Perplexity versus Sentence Counts

Preliminary Results/ 3

→ Individual sentences:

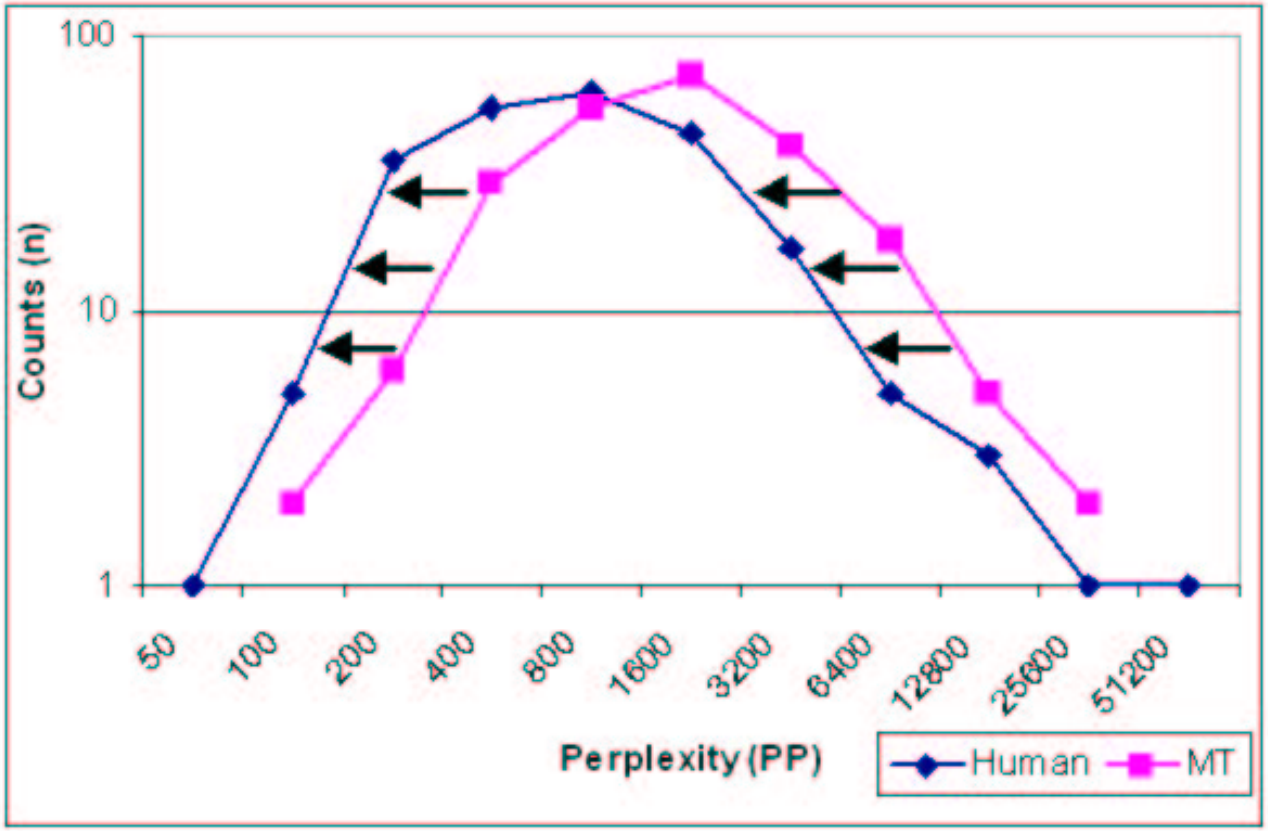


Figure 7: Perplexity versus Sentence Counts

Main Difficulties

Main Difficulties

- Sentence length influences probabilities

Main Difficulties

- Sentence length influences probabilities
- Sophisticated normalisation is necessary

Main Difficulties

- Sentence length influences probabilities
- Sophisticated normalisation is necessary
- Minimise overlap between good and bad

Summary

Summary

- Motivation for automatic error spotting: fast, cheap and desirable for users of state-of-the-art MT engines

Summary

- Motivation for automatic error spotting: fast, cheap and desirable for users of state-of-the-art MT engines

- Advantages of using a SLM: SL independent, easy to employ

Summary

- Motivation for automatic error spotting: fast, cheap and desirable for users of state-of-the-art MT engines
- Advantages of using a SLM: SL independent, easy to employ
- My approach to the task

Summary

- Motivation for automatic error spotting: fast, cheap and desirable for users of state-of-the-art MT engines
- Advantages of using a SLM: SL independent, easy to employ
- My approach to the task
- Preliminary results: text written by humans has a smaller perplexity/ entropy than machine translation output

Summary

- Motivation for automatic error spotting: fast, cheap and desirable for users of state-of-the-art MT engines
- Advantages of using a SLM: SL independent, easy to employ
- My approach to the task
- Preliminary results: text written by humans has a smaller perplexity/ entropy than machine translation output
- Anticipated difficulties: normalisation