Machine Translation Evaluation

Hervé Blanchon

Laboratoire LIG

Équipe GETALP

herve.blanchon@univ-grenoble-alpes.fr

Foreword

- What is this session about?
  - Machine translation systems evaluation
- What will we cover?
  - Evaluation by humans: subjective evaluation
    - measures
    - pros & cons of subjective evaluation
  - Efforts to formalize MT systems evaluation
  - Evaluation by programs: objective evaluation
    - measures
    - pros & cons of objective evaluation
  - Some proposals to do better

Outline

- Subjective evaluation foundations
- “Let’s try to formalize” efforts
- Subjective evaluation in practice
- Subjective evaluation final remarks
- Objective evaluation
- Objective evaluation final remarks
- Conclusion
- Bibliography

SUBJECTIVE EVALUATION FOUNDATIONS

Important dates

- 1966: ALPAC, the (In-)famous report
  - ***A**utomatic **L**anguage **P**rocessing **A**dvisory **C**ommittee*
- 1989 & 1992: JEIDA
  - ***J**apanese **E**lectronic **I**ndustry **D**evelopment **A**ssociation*
- 1992 & 1994: ARPA
  - ***A**dvanced **R**esearch **P**rojects **A**gency*
- 2000- : NIST
  - ***N**ational **I**ndustry **S**tandards and **T**echnology*

ALPAC (1966)

***A**utomatic **L**anguage **P**rocessing **A**dvisory **C**ommittee*

[ALPAC, 1966]

- An Experiment in Evaluating the Quality of Translations
  - (Appendix 10)
- Comment
  - Poor MT performance led to cuts in MT funding in the United-States
  - Highly influential work

ALPAC

- 2 major independent characteristics of a translation
  - Its intelligibility
  - Its fidelity to the sense of the original text
- Subjective rating
  - Rating of intelligibility without reference to the source
  - Indirect rating of fidelity
    - Gather whatever possible meaning from the translation sentence
    - Evaluate the source sentence “informativeness” in relation to the understanding from the translation sentence
      - A highly informative source sentence implies that the translation is lacking in fidelity

ALPAC

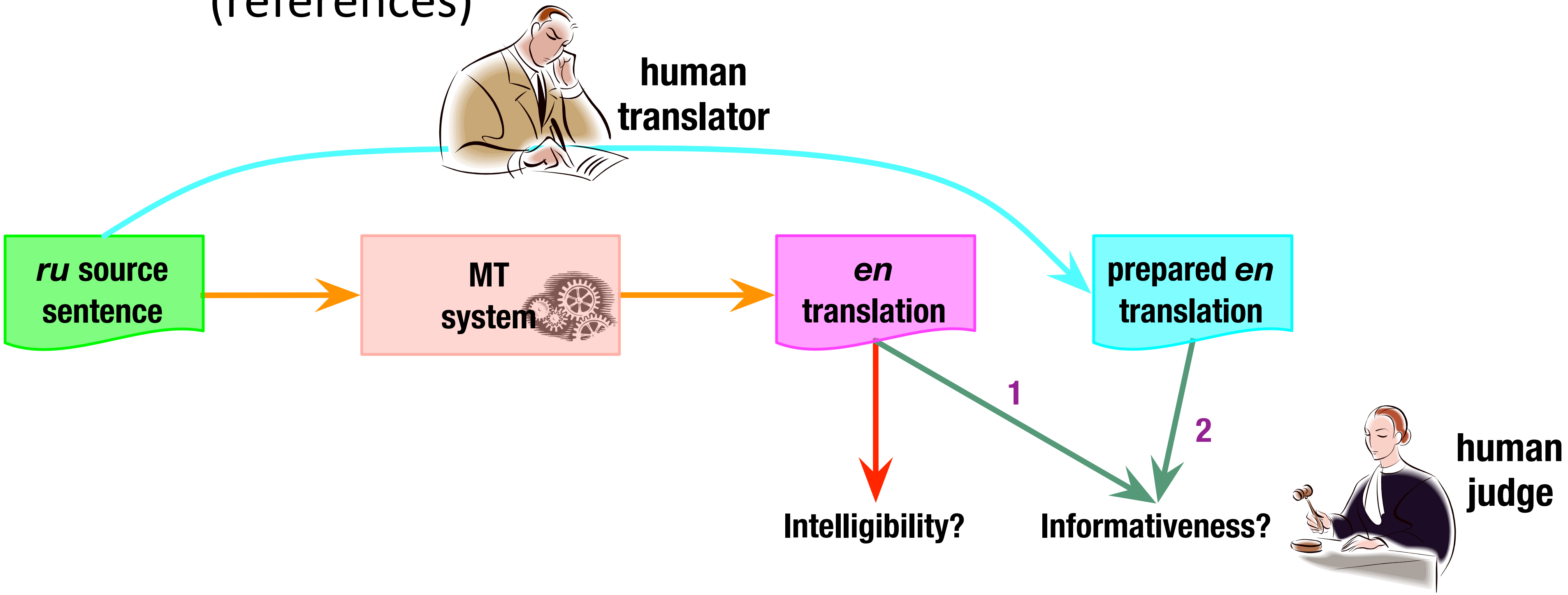
- Language pair / Domain
  - Russian → English / Scientific
- Data
  - 36 sentences / 6 translations (3 human, 3 MT systems)

ALPAC

2 sets of evaluation (1/2)

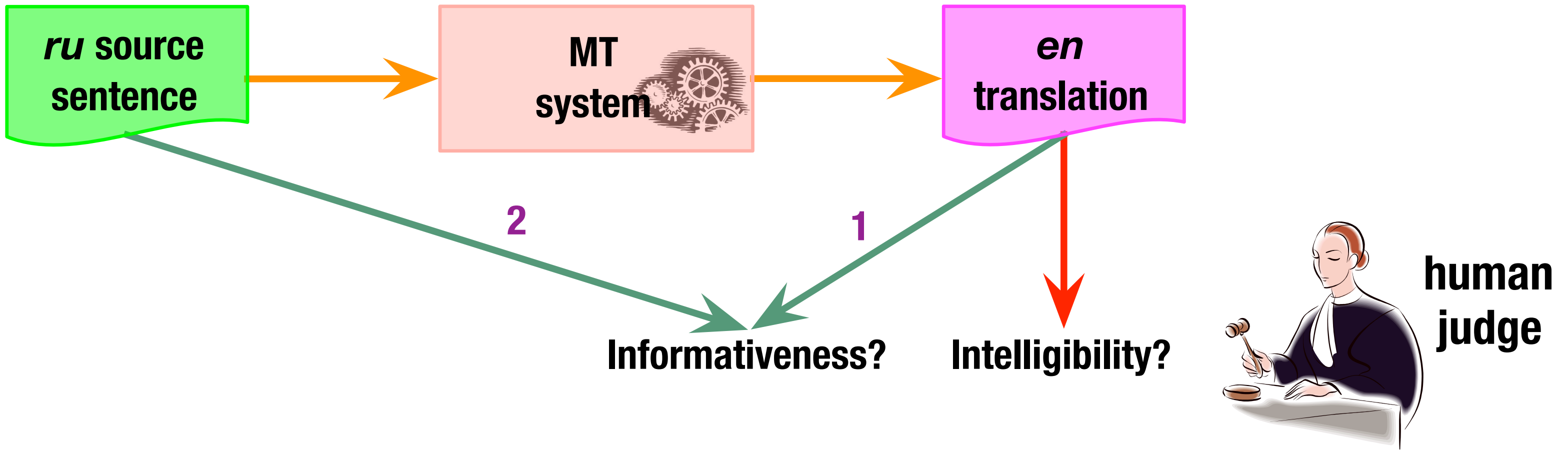
Monolingual evaluation

- 18 native English speakers with no knowledge of Russian and good background in science
- Carefully prepared English translation of the source sentences (references)



ALPAC

- 2 sets of evaluation (1/2)
  - Bilingual evaluation
    - 18 native English speakers with a high degree of competence in comprehension of scientific Russian



ALPAC: Intelligibility

9– Perfectly clear and intelligible. Reads like ordinary text; has no stylistic infelicities.

8– Perfectly or almost clear and intelligible, but contains minor grammatical or stylistic infelicities, and/or midly unusual word usage that could, nevertheless, be easily "corrected."

7– Generally clear and intelligible, but style and word choice and/or syntactical arrangement are somewhat poorer than in category 8.

6– The general idea is almost immediately intelligible, but full comprehension is distinctly interfered with by poor style, poor word choice, alternative expressions, untranslated words, and incorrect grammatical arrangements. Postediting could leave this in nearly acceptable form.

5– The general idea is intelligible only after considerable study, but after this study one is fairly confident that he understands. Poor word choice, grotesque syntactic arrangement, untranslated words, and similar phenomena are present, but constitute mainly "noise" through which the main idea is still perceptible.

4– Masquerades as an intelligible sentence, but actually it is more unintelligible than intelligible. Nevertheless, the idea can still be vaguely apprehended. Word choice, syntactic arrangement, and/or alternative expressions are generally bizarre, and there may be critical words untranslated.

3– Generally unintelligible; it tends to read like nonsense but, with a considerable amount of reflection and study, one can at least hypothesize the idea intended by the sentence.

2– Almost hopelessly unintelligible even after reflection and study. Nevertheless, it does not seem completely nonsensical.

1– Hopelessly unintelligible. It appears that no amount of study and reflection would reveal the thought of the sentence.

ALPAC: Informativeness

9– Extremely informative. Makes “all the difference in the world” in comprehending the meaning intended. (A rating of 9 should always be assigned when the original completely changes or reverses the meaning conveyed by the translation.)

8– Very informative. Contributes a great deal to the clarification of the meaning intended. By correcting sentence structure, words, and phrases, it makes a great change in the reader's impression of the meaning intended, although not so much as to change or reverse the meaning completely.

7– (Between 6 and 8.)

6– Clearly informative. Adds considerable information about the sentence structure and individual words, putting the reader “on the right track” as to the meaning intended.

5– (Between 4 and 6.)

4– In contrast to 3, adds a certain amount of information about the sentence structure and syntactical relationships; it may also correct minor misapprehensions about the general meaning of the sentence or the meaning of individual words.

3– By correcting one or two possibly critical meanings, chiefly on the word level, it gives a slightly different “twist” to the meaning conveyed by the translation. It adds no new information about sentence structure, however.

2– No really new meaning is added by the original, either at the word level or the grammatical level, but the reader is somewhat more confident that he apprehends the meaning intended.

1– Not informative at all; no new meaning is added, nor is the reader's confidence in his understanding increased or enhanced.

0– The original contains, if anything, less information than the translation. The translator has added certain meanings, apparently to make the passage more understandable.

ALPAC

- Quotes
  - "MT presumably means going by algorithm from machine-readable source text to useful target text, without recourse to human translation or editing." → "In this context, there has been no machine translation of general scientific text, and none is in immediate prospect."
  - "The reader will find it instructive to compare the samples above with the results obtained on simple, selected, text 10 years earlier (the Georgetown IBM Experiment, January 7, 1954) in that the earlier samples are more readable than the later ones."
  - In the final chapter (p.32-33), ALPAC underlined once more that "we do not have useful machine translation [and] there is no immediate or predictable prospect of useful machine translation." It repeated the potential opportunities to improve translation quality, particularly in various machine aids: "Machine-aided translation may be an important avenue toward better, quicker, and cheaper translation." But ALPAC did not recommend basic research: "What machine-aided translation needs most is good engineering."

Jeida (1989 & 1992)

Japanese ***E****lectronic* ***I****ndustry* ***D****evelopment* ***A****ssociation*

- Jeida 1989 [JEIDA, 1989]
  - *A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC.*
  - 3 questions
    - What are the technological and social changes of the market since the ALPAC report?
    - According to these changes, are the conclusions of the ALPAC report still valid today?
    - If not, how should we evaluate the current state and the future of machine translation?
  - No clear answer!

Jeida (1989 & 1992)

- Jeida 1992 [JEIDA, 1992]
  - JEIDA Methodology and Criteria on Machine Translation Evaluation
  - Several point of view using complex forms
    - Economical factors evaluation by the users
    - Technical evaluation of the systems by the users
      - “Satisfaction of the users’ needs”
    - Technical evaluation of the systems by the developers
      - “Criteria to help researchers, developers, and project leaders in evaluating their systems”

ARPA (1992-1994) & NIST (2000-)

***A**dvanced **R**esearch **P**rojects **A**gency*

***N**ational **I**ndustry **S**tandards and **T**echnology*

- Comparative/competitive evaluation [White et al, 1994]
  - Systems
    - Fully automatic / Human Aided MT
  - Language pairs
    - Source language: several / Target language: English
  - Domain
    - Newspaper articles about financial mergers and acquisitions
    - *Professionally translated into the respective source languages or into English*
  - Evaluators
    - literate, monolingual English speakers

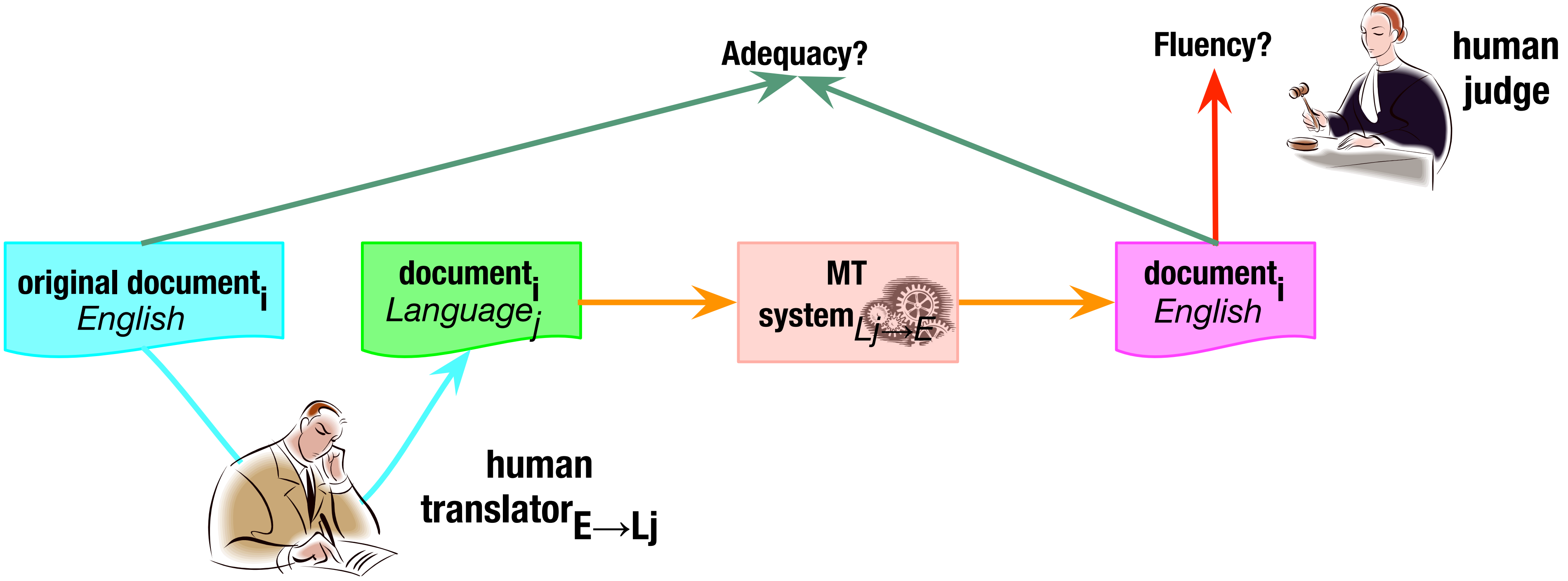
ARPA & NIST

- Criteria
  - Fluency
    - without reference to the source
  - Adequacy
    - in contrast to the the English original or translation

Score	Adequacy	Fluency
5	All information	Flawless English
4	Most	Good
3	Much	Non-Native
2	Little	Disfluent
1	None	Incomprehensible

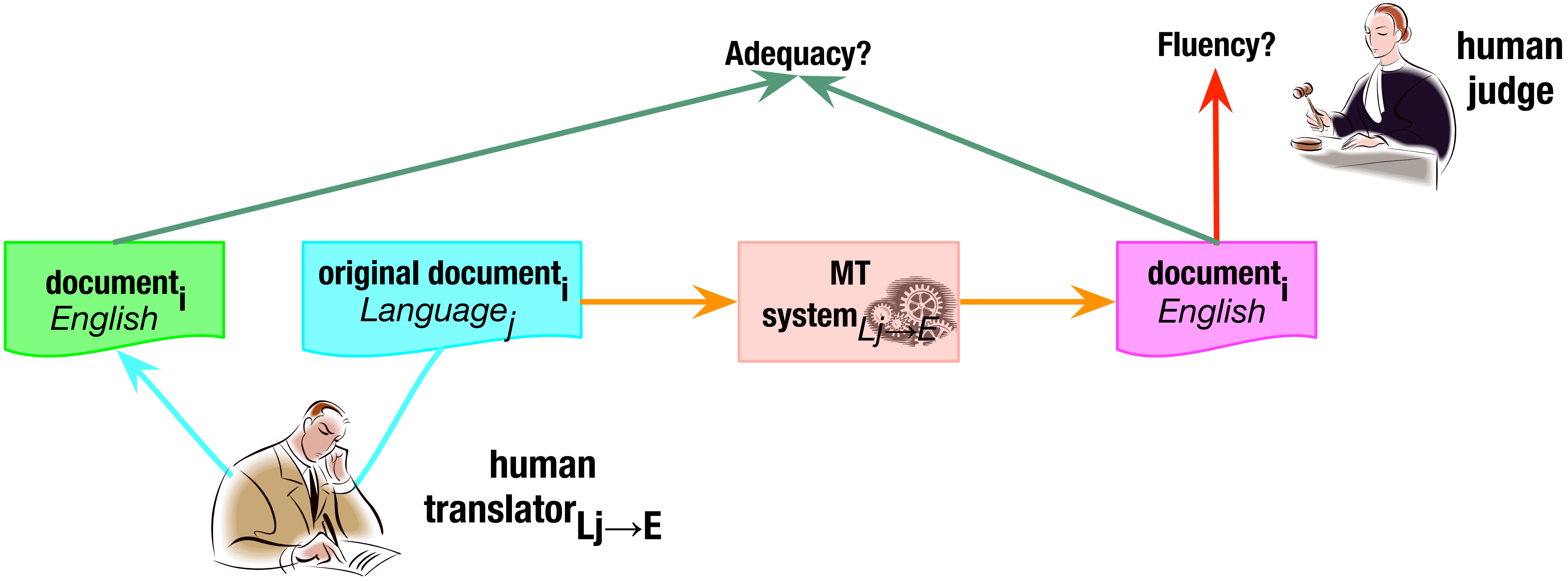
ARPA & NIST

- When source document is not available



ARPA & NIST

- When source document is available


Adequacy?
Fluency?
human judge
document$_i$ *English*
original document$_i$ *Language$_j$*
MT system$_{Lj\rightarrow E}$
document$_i$ *English*
human translator$_{Lj\rightarrow E}$

“LET’S TRY TO FORMALIZE” EFFORTS

Important dates

- 1993-1996: EAGLES
  - ***E**xpert **A**dvisory **G**roup on **L**anguage **E**ngineering*
  - Initiative of the European Commission
  - [EAGLES-EWG, 1996] [EAGLES-EWG, 1999]
- 1999-2002: ISLE (FEMTI)
  - ***F**ram**e**work for **M**achine **T**ranslation Evaluation on **I**SLE (International Standards for Language Engineering)*
  - Joined initiative of the European Commission and National Science Foundation (NSF)
  - http://www.isi.edu/natural-language/mteval/
  - [Hovy et al., 2002] [King et al., 2003]

EAGLES

***E**xpert **A**dvisory **G**roup on **L**anguage **E**ngineering*

- Goal
  - Standards for the language engineering industry
- Targets
  - Corpora
  - Lexicons
  - Grammatical formalisms
  - Evaluation
- On evaluation
  - A quality model for natural language processing tools…
  - … validated on grammar checkers,

EAGLES: A 7-steps Recipe

1. Why is the evaluation being done?
2. Elaborate a task model
3. Define top level quality characteristics
4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3
5. Devise the metrics to be applied to the system for the requirements produced under 4
6. Design the execution of the evaluation
7. Execute the evaluation

EAGLES: A 7-steps Recipe

1. Why is the evaluation being done?

- What is the purpose of the evaluation? Do all parties involved have the same understanding of the purpose?
- What exactly is being evaluated? Is it a system or a system component? A system in isolation or a system in a specific context of use? Where are the boundaries of the system?

2. Elaborate a task model

- Identify all relevant roles and agents
- What is the system going to be used for?
- Who will use it? What will they do with it? What are these people like?

3. Define top level quality characteristics

- What features of the system need to be evaluated? Are they all equally important?

EAGLES: A 7-steps Recipe

4. Produce detailed requirements for the system under evaluation, on the basis of 2 and 3

- For each feature which has been identified as important, can a valid and reliable way be found of measuring how the object being evaluated performs with respect to that feature?
- If not, then the features have to be broken down in a valid way, into sub-attributes which are measurable.
- This point has to be repeated until a point is reached where the attributes are measurable.

EAGLES: A 7-steps Recipe

5. Devise the metrics to be applied to the system for the requirements produced under 4

- Both measure and method for obtaining that measure have to be defined for each attribute.
- For each measurable attribute, what will count as a good score, a satisfactory score or an unsatisfactory score given the task model (2)? Where are the cut off points?
- Usually, an attribute has more than one sub-attributes. How are the values of the different sub-attributes combined to a value for the mother node in order to reflect their relative importance (again given the task model)?

EAGLES: A 7-steps Recipe

6. Design the execution of the evaluation

- Develop test materials to support the testing of the object.
- Who will actually carry out the different measurements? When? In what circumstances? What form will the end result take?

7. Execute the evaluation:

- Make measurement.
- Compare with the previously determined satisfaction ratings.
- Summarize the results in an evaluation report, cf. point 1.

FEMTI

F*ram***e***work for* ***M****achine* ***T****ranslation Evaluation on* ***I****SLE (International Standards for Language Engineering)*

- Attempt to organize the various methods for MT evaluation

FEMTI

- FEMTI contains
  - A classification of the main features defining the context of use (type of user of the MT system, type of task the system is used for, nature of the input to the system)
  - A classification of the MT software quality characteristics, into hierarchies of sub-characteristics, with internal and/or external attributes (i.e., metrics) at the bottom level.
  - A mapping from the first classification to the second, which defines or suggests the quality characteristics, sub-characteristics and attributes/metrics that are relevant to each context of use.

FEMTI (top level classification)

1 Evaluation requirements

- 1.1 The purpose of evaluation
- 1.2 The object of evaluation
- 1.3 Characteristics of the translation task
  - 1.3.1 Assimilation
  - 1.3.2 Dissemination
  - 1.3.3 Communication
- 1.4 User characteristics
  - 1.4.1 Machine translation user
  - 1.4.2 Translation consumer
  - 1.4.3 Organisational user
- 1.5 Input characteristics (author and text)
  - 1.5.1 Document type (genre, domain/field of application)
  - 1.5.2 Author characteristics (proficiency in source language, training)
  - 1.5.3 Characteristics related to sources of errors (unproofed text)

FEMTI (top level classification)

2 System characteristics to be evaluated

- 2.1 System internal characteristics
  - 2.1.1 MT system-specific characteristics
  - 2.1.2 Translation process models
  - 2.1.3 Linguistic resources and utilities
  - 2.1.4 Characteristics of process flow
- 2.2 System external characteristics
  - 2.2.1 Functionality
    - 2.2.1.1 Suitability, Accuracy, Wellformedness, Interoperability, Compliance, Security
  - 2.2.2 Reliability
  - 2.2.3 Usability
  - 2.2.4 Efficiency
  - 2.2.5 Maintainability
  - 2.2.6 Portability
  - 2.2.7 Cost

FEMTI (Section 2.2.1 Functionality)

- 2.2.1.1 Suitability
  - 2.2.1.1.1 Target-language only
    - 2.2.1.1.1.1 Readability (or: fluency, intelligibility, clarity)
    - 2.2.1.1.1.2 Comprehensibility
    - 2.2.1.1.1.3 Coherence
    - 2.2.1.1.1.4 Cohesion
  - 2.2.1.1.2 Cross-language / contrastive
    - 2.2.1.1.2.1 Coverage of corpus-specific phenomena
    - 2.2.1.1.2.2 Style
- 2.2.1.2 Accuracy
  - 2.2.1.2.1 Fidelity
  - 2.2.1.2.2 Consistency
  - 2.2.1.2.3 Terminology

FEMTI (Section 2.2.1 Functionality [cont.])

- 2.2.1.3 Wellformedness
  - 2.2.1.3.1 Punctuation
  - 2.2.1.3.2 Lexis / lexical choice
  - 2.2.1.3.3 Grammar / syntax
  - 2.2.1.3.4 Morphology
- 2.2.1.4 Interoperability
- 2.2.1.5 Compliance
- 2.2.1.6 Security

FEMTI (2.2.1.1.1.1 readability)

- Definition
  - The extent to which a sentence reads naturally.
  - Ease with which a translation can be understood, i.e. its clarity to the reader. (Halliday in Van Slype's Critical Report) .
  - This has also been called fluency, intelligibility, and clarity.
- Metrics
  - …
  - Pfafflin (in Van Slype's Critical Report): Rating of sentences read on a 3-point scale.
  - Vanni & Miller (2001, 2002): "Do you get it?" - snap judgement rating of sentences on scale from 0 to 3.
  - Niessen, Och, Leusch and Ney, 2000 measure syntactic errors with an automated string edit distance metric, which according to them can also be used as a measure of readability. See also Wellformedness (2.2.1.3/186).
  - J.B. Carroll: by measuring the time spent by the evaluator in reading each sentence of the sample.
  - Pfafflin and Orr (both quoted by T.C. Halliday): by measuring the response time to a multiple-choice questionnaire.
  - H.W. Sinaiko: by measuring the time necessary for the execution of the cloze test.
- Notes
  - Readability is intended to be a metric applied at the sentence-level. …
  - Readability is a quality of the output that can be measured independently of the source language.
  - Cloze tests can be used either at sentence-level or cross-sentence level.
  - This quality has been merged with clarity, which was a separate taxon in earlier versions of this taxonomy.

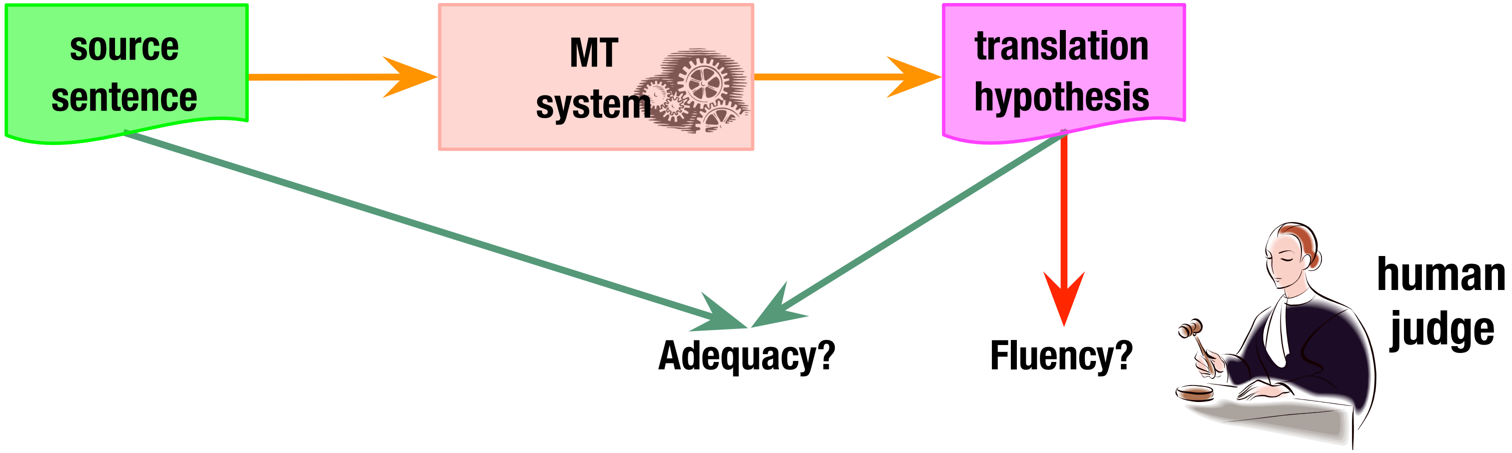
FEMTI (2.2.1.2.1 Fidelity)

- Definition
  - Subjective evaluation of the degree to which the information contained in the original text has been reproduced without distortion in the translation (Van Slype).
  - Measurement of the correctness of the information transferred from the source language to the target language (Halliday in Van Slype's Critical Report).
- Metrics
  - …
  - White and O'Connell (in DARPA 94): Rating of 'Adequacy' on a 5-point scale.
  - Bleu evaluation tool kit (in Papineni et al. 2001): Automatic n-gram comparison of translated sentences with one or more human reference translations.
  - Rank-order evaluation of MT system: correlation of automatically computed semantic and syntactic attributes of the MT output with human scores for adequacy and informativeness, and also fluency. Hartley and Rajman 2001 and 2002.
  - Automated word-error-rate evaluation (in Och, Tillmann and Ney, 1999).
- Notes
  - The fidelity rating has been found to be equal to or lower than the comprehensibility rating, since the unintelligible part of the message is not found in the translation. Any variation between the comprehensibility rating and the fidelity rating is due to additional distortion of the information, which can arise from: – loss of information (silence) - example: word not translated, – interference (noise) - example: word added by the system, – distortion from a combination of loss and interference - example: word badly translated.
  - Detailed analysis of the fidelity of a translation is very difficult to carry out, since each sentence conveys not a single item of information or a series of elementary items of information, but rather a portion of message or a series of complex messages whose relative importance in the sentence is not easy to appreciate.

SUBJECTIVE EVALUATION IN PRACTICE

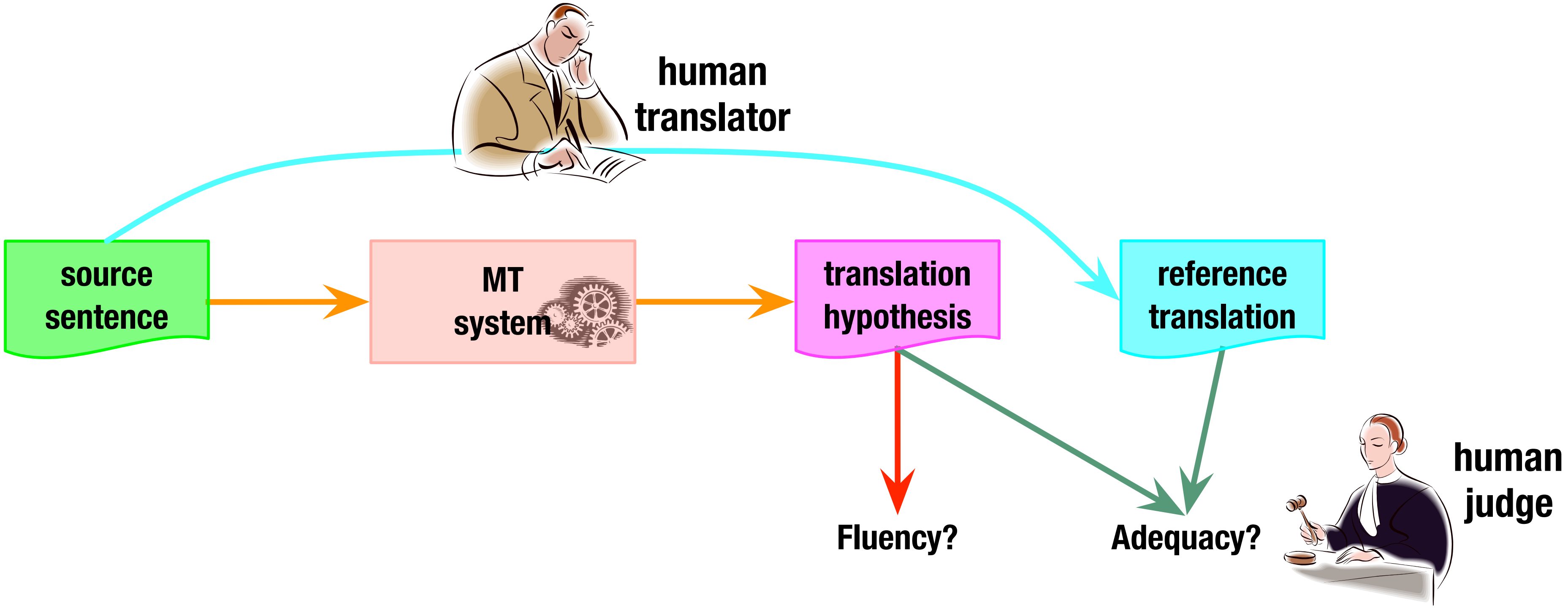
The settings

- Bilingual evaluation



The settings

- Monolingual evaluation



Example: NESPOLE! (2002)

- Language pairs
  - French → Italian
  - Italian → French
  - French → French (speech recognition results)
- Domain
  - Tourism
- Evaluators
  - Trained students from a translation school
  - 3 Native French speakers for Italian → French & French → French
  - 3 Native Italian speakers for French → Italian
- Evaluation criterion
  - Quality of the translation on a 4 grades scale
    - **Very good** (all information & easy to understand), **Good** (all important information)
    - **Bad** (one or several important information missing), **Very Bad** (almost all important information missing)

Example: NESPOLE! (2002)

- Protocol
  - Trained evaluators
  - Evaluators inter-agreement
    - Same data evaluated by each group (French, Italian): control test set
  - Evaluators self-agreement
    - Same data evaluated by each evaluator before & after the actual evaluation: control test set

Example: NESPOLE! (2002)

- Evaluators inter-agreement
  - French or Italian → French

	Unanimity	Majority	No Majority
Before the task	71 %	28 %	1 %
After the task	73 %	27 %	0 %

  - French → Italian

	Unanimity	Majority	No Majority
Before the task	88 %	15 %	0 %
After the task	75 %	25 %	0 %

- Conclusion
  - Fairly good inter-agreement

Example: NESPOLE! (2002)

- Evaluators self-agreement (before vs after, over 102)
  - French evaluators

	=	class =	class ≠
Eval1	58	27	17
Eval2	83	13	6
Eval3	65	18	19

  - Italian evaluators

	=	class =	class ≠
Eval4	83	13	6
Eval5	102	0	0
Eval6	58	27	17

- Conclusion
  - Evaluators tend to be more harsh, scores are always lowered **VG** to **G**, **B** to **VB** or (**VG**,**G**) to (**B**, **VB**)

Very **G**ood, **G**ood, **B**ad, **V**ery **B**ad

Example: NESPOLE! (2002)

Evaluation Excel file (It → Fr)

			Very Good	Good	Bad	Very Bad
TURN	*1*		-----	-----	-----	-----
1	APT del trentino	1	x			
2	buongiorno	2	x			
TRANS	Ici une agence d'informations du Trentin. Bonjour !	-	-----	-----	-----	-----
TURN	*39*	-	-----	-----	-----	-----
1	sì	1	x			
2	poi ci sono incluse nel pacchetto 4 lezioni di sci e 2 lezioni pattinaggio	2	x			
TRANS	Oui. Il y a un forfait avec 4 leçons du ski. Le 2 du patin.	-	-----	-----	-----	-----

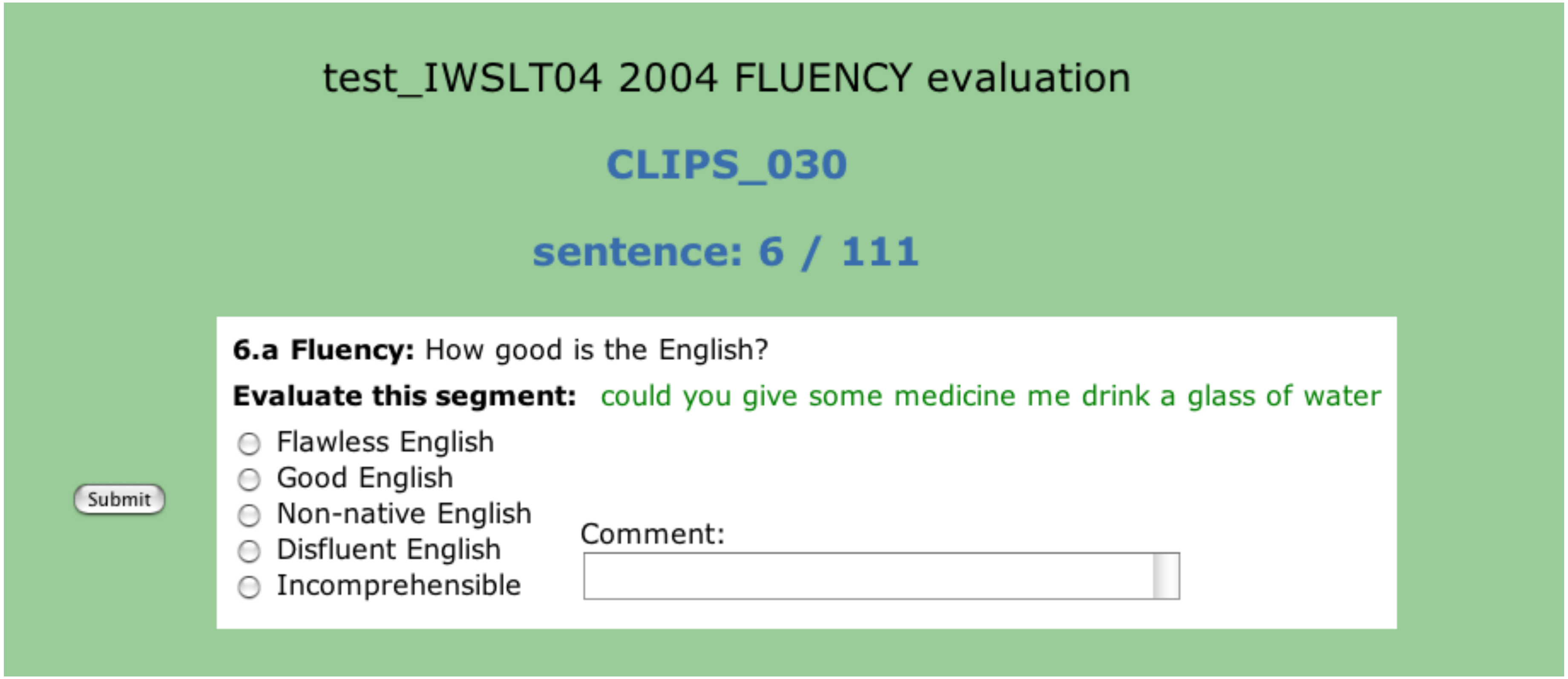
- ✓APT del trentino = trentino tourism agency
- ✓Buongiorno = good moring, hello
- ✓Ici une agence d'information du Trentin. Bonjour ! = Here an information agency of Trentino. Good Morning!
- ✓Sì = yes
- ✓poi ci sono incluse nel pacchetto 4 lezioni di sci e 2 lezioni pattinaggio = then are included in the package 4 ski lessons and 2 skating lessons
- ✓Oui. Il y a un forfait avec 4 leçons du ski. Le 2 du patin. = Yes. There is a package with 4 lessons of the ski. The 2 of skating.

Example: IWSLT (2004)

- Language pair
  - Japanese → English
- Domain
  - Tourism
- Evaluators
  - Native English speakers
- Evaluation criterion
  - Fluency
  - Adequacy

Example: IWSLT (2004)

- Fluency



test_IWSLT04 2004 FLUENCY evaluation

CLIPS_030

sentence: 6 / 111

6.a Fluency: How good is the English?

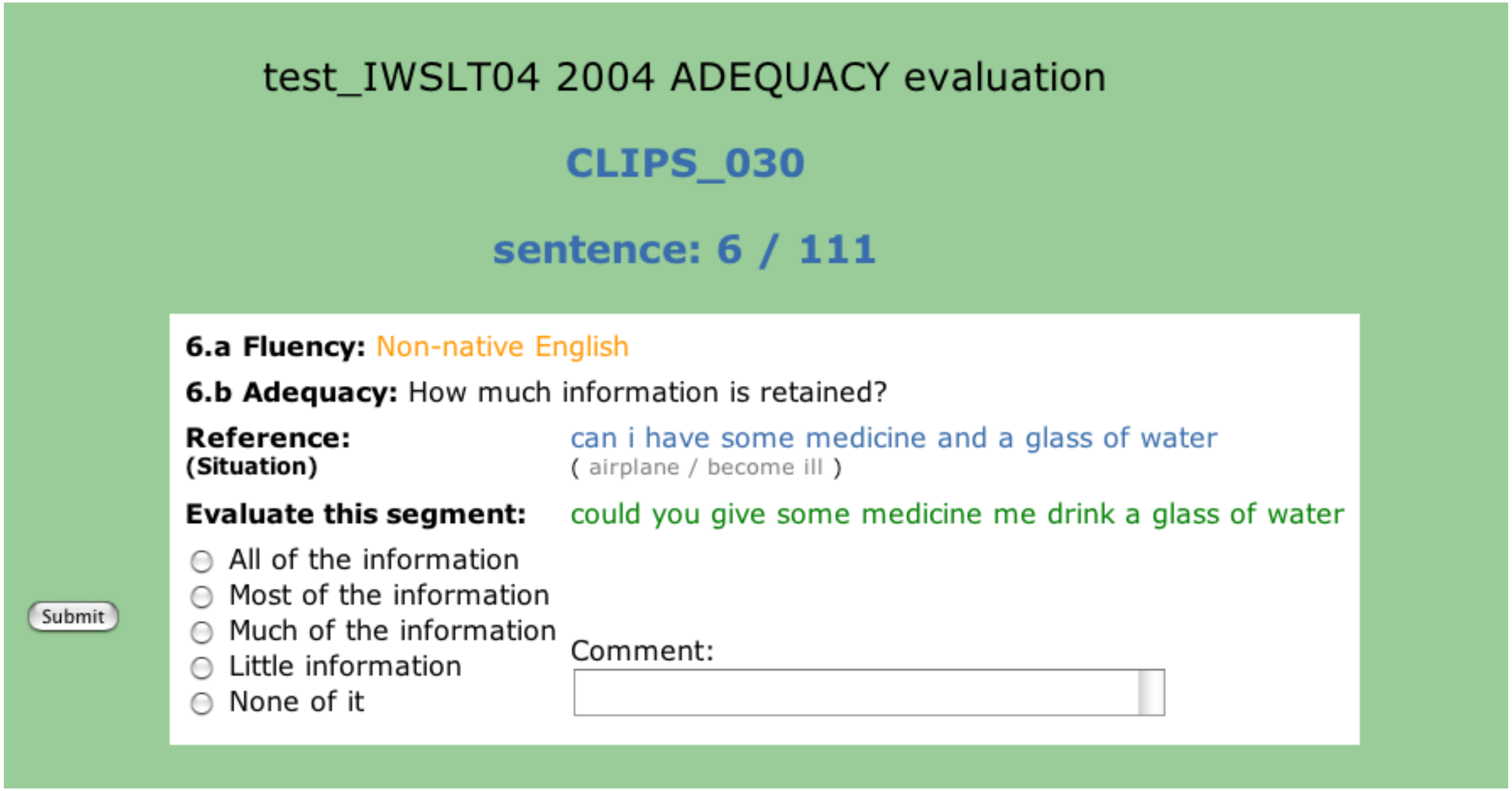
Evaluate this segment: could you give some medicine me drink a glass of water

○ Flawless English
○ Good English
○ Non-native English
○ Disfluent English
○ Incomprehensible

Comment:

Example: IWSLT (2004)

- Adequacy



test_IWSLT04 2004 ADEQUACY evaluation

CLIPS_030

sentence: 6 / 111

6.a Fluency: Non-native English

6.b Adequacy: How much information is retained?

Reference: can i have some medicine and a glass of water
(Situation) (airplane / become ill)

Evaluate this segment: could you give some medicine me drink a glass of water

○ All of the information
○ Most of the information
○ Much of the information
○ Little information
○ None of it

Comment:

SUBJECTIVE EVALUATION

FINAL REMARKS

Pro of subjective evaluation

- Very informative

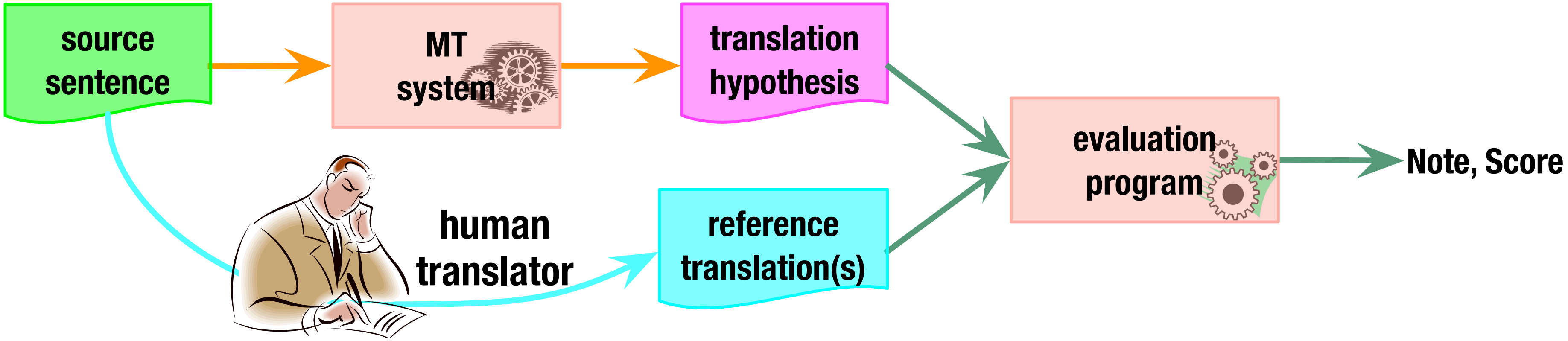
Cons of subjective evaluation

- Labor-intensive & Time-consuming (Evaluators, Translators)
  - In practice, impossible for evaluation campaigns (subset or one run evaluation organized as a shared task between participants)
- Not reusable
  - MT systems as dynamic components improving along time
  - Human assessment as a one shot measure to be repeated
- Subjective
  - Evaluators' understanding of the guidelines
  - Evaluators' inter-agreement
  - Evaluators' intra-agreement
- Possibly partial
  - Mostly limited to fluency and adequacy
  - Difficulty to compare
    - E.g. fluency(SystA)<fluency(SystB) & adequacy(SystA)>adaquacy(SystB) …
    - … Best(SystA, SystB) or Best(SystB, SystA)??????

OBJECTIVE EVALUATION

Ideas

- Get ride of …
  - Subjectivity, Non reusability, Slowness, Expensiveness
- How?
  - Take advantage of the reference(s) produced for subjective evaluation
  - Use a deterministic program to compare hypothesis with reference(s)



Important dates

- 2002: BLEU [Papineni et al. 2002]
  - The beginning of objective evaluation measures
- Systems evaluation campaigns
  - 2001-: NIST Open MT
    - http://www.itl.nist.gov/iad/mig/tests/mt/
  - 2004-: IWSLT
    - Speech translation
    - http://iwslt2011.org/doku.php?id=14_related_events
  - 2006- : WMT
    - Broadcast news
    - http://www.statmt.org/wmt12/
- Metrics evaluation campaigns
  - 2008-: NIST MetricsMaTr
    - Metrics for Machine Translation Evaluation
    - http://www.nist.gov/itl/iad/mig/metricsmatr.cfm

Summary

- The rough idea: lexical similarity
- Several measures*
  - Edit distance measures
    - WER, PER, TER
  - Precision-oriented measures
    - BLEU, NIST, WNM
  - Recall-oriented measures
    - ROUGE, CDER
  - Balancing precision & recall measures
    - GTM, METEOR, BLANC, SIA

*Incomplete because new measures are proposed every other day!!

Edit distance measures

Number of changes:

hypothesis → reference or acceptable translation

- **WER** (Word Error Rate) [Nießen et al., 2000]
  - Based on the Leveinstein distance: minimum number of substitutions, deletions, or insertions that have to be performed to convert the hypothesis into the reference
- **PER** (Position-independent Word Error Rate) [Tillmann et al., 1997]
  - A shortcoming of WER, PER compare the words in the hypothesis and reference without taking into account word order (bags of words)
- **TER** (Translation Edit Rate) [Snover et al. 2006] [Przybocki et al. 2006]
  - Operations performed by a post-editor to correct the hypothesis (insertion, deletion, substitution of words or sequences)

WER

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Translation 2: the green house was by the lake shore .
- Translation 3: the green potato right in front of the lake was right .
- Translation 4: the green house was right in front of the lake .

	WER
T1	54.5455
T2	45.4545
T3	36.3636
T4	00.0000

WER

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .

- Computation

```
REF:  the green house was right in front of the lake ***** .
HYP:  a   green house was ***** ** ***** by the lake shore .
EVAL: S                   D     D  D     S           I
SHFT:
WER Score:  54,55 (  6,0/ 11,0)
```

WER

- Reference: the green house was right in front of the lake .
- Translation 1: the green house was by the lake shore .

- Computation

```
REF:  the green house was right in front of the lake ***** .
HYP:  the green house was ***** ** ***** by the lake shore .
EVAL:                     D     D  D     S           I
SHFT:
WER Score:  45,45 (  5,0/ 11,0)
```

WER

- Reference: the green house was right in front of the lake .
- Translation 1: the green potato right in front of the lake was right .

- Computation

```
REF:   the green house was    right in front of the lake *** ***** .
HYP:   the green ***** potato right in front of the lake was right .
EVAL:              D     S                                I   I
SHFT:
TER Score:  36,36 (  4,0/ 11,0)
```

PER

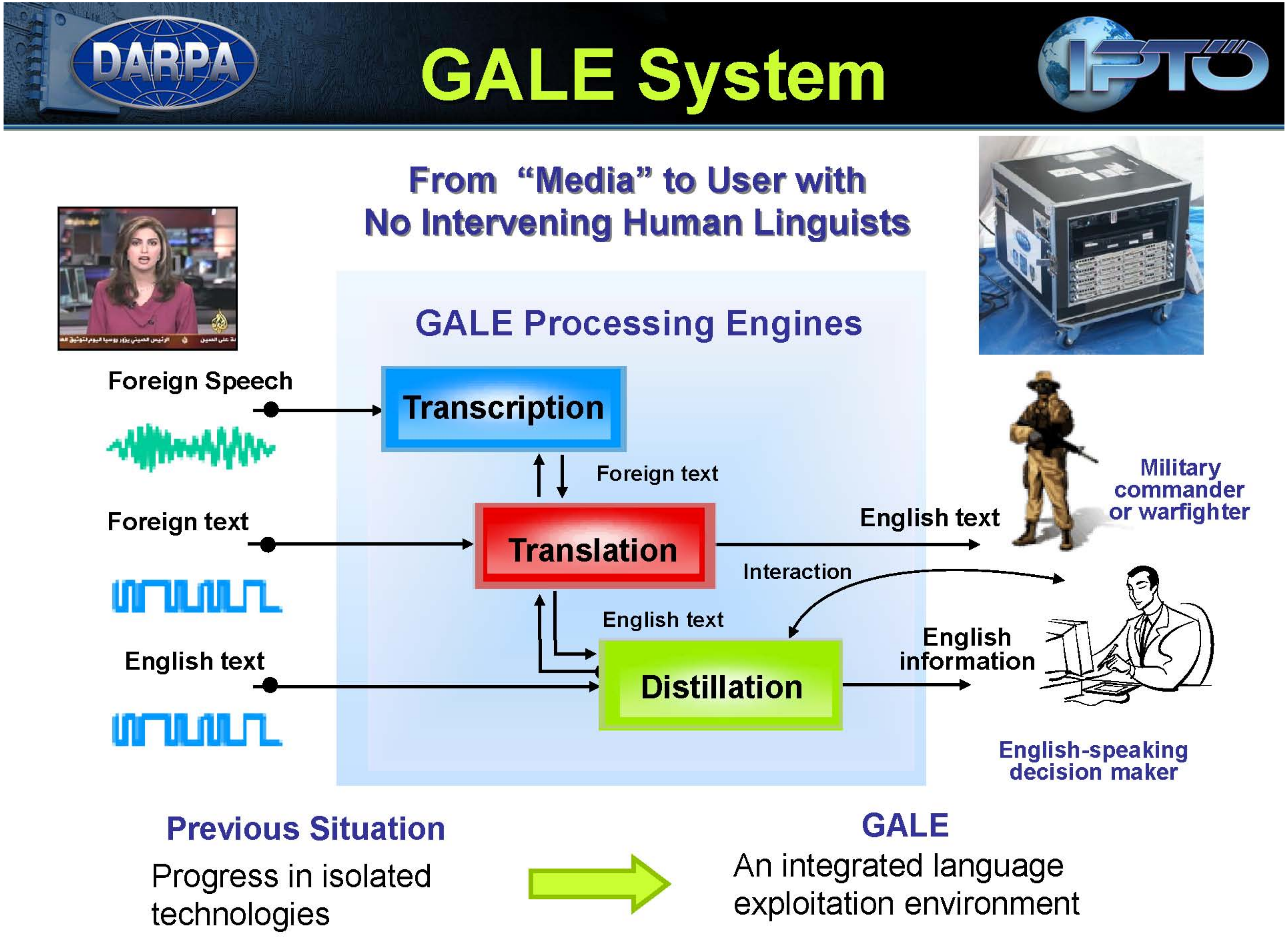
- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Translation 2: the green house was by the lake shore .
- Translation 3: the green potato right in front of the lake was right .
- Translation 4: the green house was right in front of the lake.

	PER
T1	45.4545
T2	36.3636
T3	18.1818
T4	00.0000

TER in GALE (HTER)

- GALE (global autonomous language exploitation) program (DARPA, 05-06)
  - develop and apply computer software technologies to absorb, translate, analyze, and interpret huge volumes of speech and text in multiple languages
  - evaluation for “go, no-go” funding
  - http://www.darpa.mil/Our_Work/I2O/Programs/Global_Autonomous_Language_Exploitation_(GALE).aspx

GALE



GALE System

From "Media" to User with
No Intervening Human Linguists

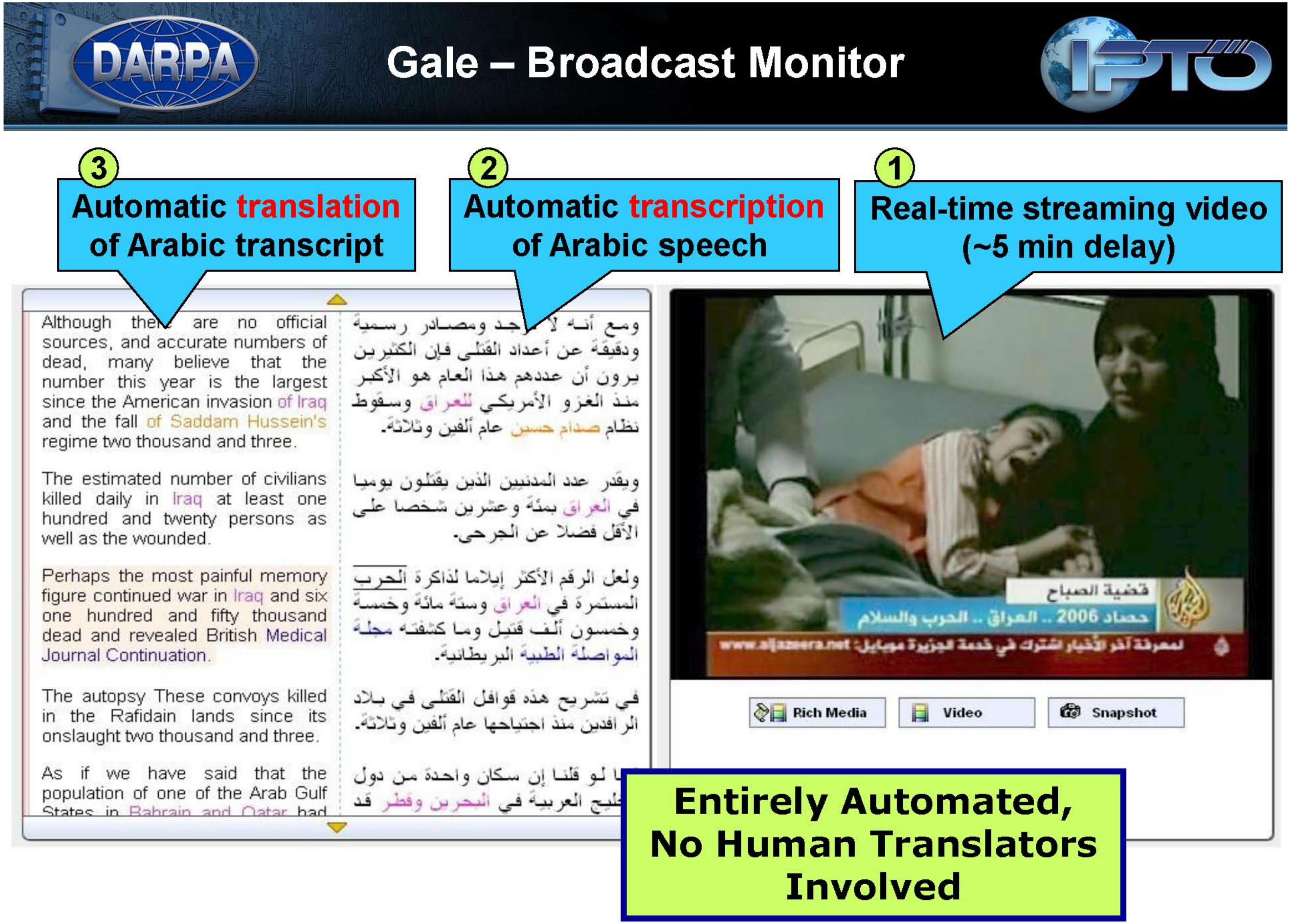
Previous Situation

Progress in isolated technologies

GALE

An integrated language exploitation environment

GALE


DARPA
IPTO
قضية الصباح
حصاد 2006 .. العراق .. الحرب والسلام
www.aljazeera.net لمعرفة آخر الأخبار اشترك في خدمة الجزيرة موبايل
Rich Media
Video
Snapshot

Gale – Broadcast Monitor

3 Automatic translation of Arabic transcript

2 Automatic transcription of Arabic speech

1 Real-time streaming video (~5 min delay)

Although there are no official sources, and accurate numbers of dead, many believe that the number this year is the largest since the American invasion of Iraq and the fall of Saddam Hussein's regime two thousand and three.

The estimated number of civilians killed daily in Iraq at least one hundred and twenty persons as well as the wounded.

Perhaps the most painful memory figure continued war in Iraq and six one hundred and fifty thousand dead and revealed British Medical Journal Continuation.

The autopsy These convoys killed in the Rafidain lands since its onslaught two thousand and three.

As if we have said that the population of one of the Arab Gulf States in Bahrain and Qatar had

ومع أنه لا [illegible]جد ومصادر رسمية ودقيقة عن أعداد القتلى فإن الكثيرين يرون أن عددهم هذا العام هو الأكبر منذ الغزو الأمريكي للعراق وسقوط نظام صدام حسين عام ألفين وثلاثة.

ويقدر عدد المدنيين الذين يقتلون يوميا في العراق بمئة وعشرين شخصا على الأقل فضلا عن الجرحى.

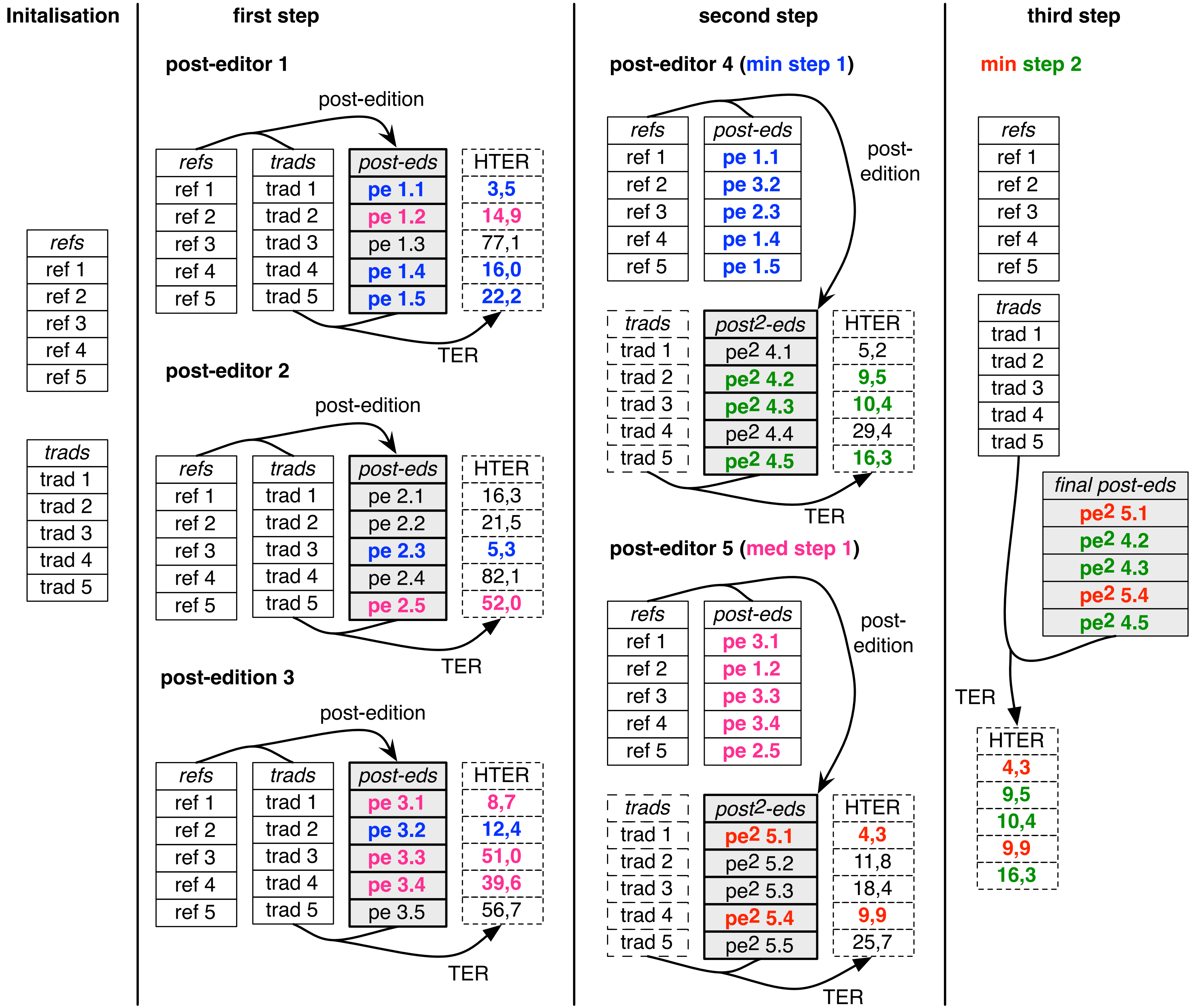
ولعل الرقم الأكثر إيلاما لذاكرة الحرب المستمرة في العراق وستة مائة وخمسة وخمسون ألف قتيل وما كشفته مجلة المواصلة الطبية البريطانية.

في تشريح هذه قوافل القتلى في بلاد الرافدين منذ اجتياحها عام ألفين وثلاثة.

[illegible]ا لو قلنا إن سكان واحدة من دول [illegible]ليج العربية في البحرين وقطر قد

Entirely Automated, No Human Translators Involved

HTER



Initalisation

refs
ref 1
ref 2
ref 3
ref 4
ref 5

trads
trad 1
trad 2
trad 3
trad 4
trad 5

first step

post-editor 1

refs	trads	post-eds	HTER
ref 1	trad 1	pe 1.1	3,5
ref 2	trad 2	pe 1.2	14,9
ref 3	trad 3	pe 1.3	77,1
ref 4	trad 4	pe 1.4	16,0
ref 5	trad 5	pe 1.5	22,2

post-editor 2

refs	trads	post-eds	HTER
ref 1	trad 1	pe 2.1	16,3
ref 2	trad 2	pe 2.2	21,5
ref 3	trad 3	pe 2.3	5,3
ref 4	trad 4	pe 2.4	82,1
ref 5	trad 5	pe 2.5	52,0

post-edition 3

refs	trads	post-eds	HTER
ref 1	trad 1	pe 3.1	8,7
ref 2	trad 2	pe 3.2	12,4
ref 3	trad 3	pe 3.3	51,0
ref 4	trad 4	pe 3.4	39,6
ref 5	trad 5	pe 3.5	56,7

second step

post-editor 4 (min step 1)

refs	post-eds
ref 1	pe 1.1
ref 2	pe 3.2
ref 3	pe 2.3
ref 4	pe 1.4
ref 5	pe 1.5

trads	post2-eds	HTER
trad 1	pe2 4.1	5,2
trad 2	pe2 4.2	9,5
trad 3	pe2 4.3	10,4
trad 4	pe2 4.4	29,4
trad 5	pe2 4.5	16,3

post-editor 5 (med step 1)

refs	post-eds
ref 1	pe 3.1
ref 2	pe 1.2
ref 3	pe 3.3
ref 4	pe 3.4
ref 5	pe 2.5

trads	post2-eds	HTER
trad 1	pe2 5.1	4,3
trad 2	pe2 5.2	11,8
trad 3	pe2 5.3	18,4
trad 4	pe2 5.4	9,9
trad 5	pe2 5.5	25,7

third step

min step 2

refs
ref 1
ref 2
ref 3
ref 4
ref 5

trads
trad 1
trad 2
trad 3
trad 4
trad 5

final post-eds
pe2 5.1
pe2 4.2
pe2 4.3
pe2 5.4
pe2 4.5

HTER
4,3
9,5
10,4
9,9
16,3

TER: examples

- Source: a burglar broke into my room .
- Best Ref: un cambrioleur a forcé ma chambre .
- Orig Hyp: un cambrioleur est entré de force dans ma pièce .

```
REF:  un cambrioleur *** ****** ** a     forcé ma chambre .
HYP:  un cambrioleur est entré  de force dans  ma pièce   .
EVAL:                I   I      I  S     S         S
SHFT:
```

✓ TER Score: 85,71 (6,0/ 7,0)

- Source: a man snatched my bag on the street .
- Best Ref: un homme a saisi mon sac dans la rue .
- Orig Hyp: un homme a saisi mon sac sur la rue .

```
REF:  un homme a saisi mon sac dans la rue .
HYP:  un homme a saisi mon sac sur  la rue .
EVAL:                          S
SHFT:
```

✓ TER Score: 10,00 (1,0/ 10,0)

TER: examples

- Source: a pickpocket took my wallet .
- Best Ref: un pickpocket a pris mon portefeuille .
- Orig Hyp: un pickpocket a pris mon portefeuille .

```
REF:  un pickpocket a pris mon portefeuille .
HYP:  un pickpocket a pris mon portefeuille .
EVAL:
SHFT:
```

✓ TER Score: 0,00 (0,0/ 7,0)

- Source: about how much would a taxi be from here .
- Best Ref: combien est-ce qu'un taxi coûterait d'ici ?
- Orig Hyp: au sujet de combien est-ce qu'un taxi serait d'ici ?

```
REF:  ** ***** ** combien est-ce qu'un taxi coûterait d'ici ?
HYP:  au sujet de combien est-ce qu'un taxi serait    d'ici ?
EVAL: I  I     I                            S
SHFT:
```

✓ TER Score: 57,14 (4,0/ 7,0)

TER: examples

- Source: about ten minutes .
- Best Ref: approximativement dix minutes .
- Orig Hyp: approximativement dix minutes .

```
REF:  approximativement dix minutes .
HYP:  approximativement dix minutes .
EVAL:
SHFT:
```

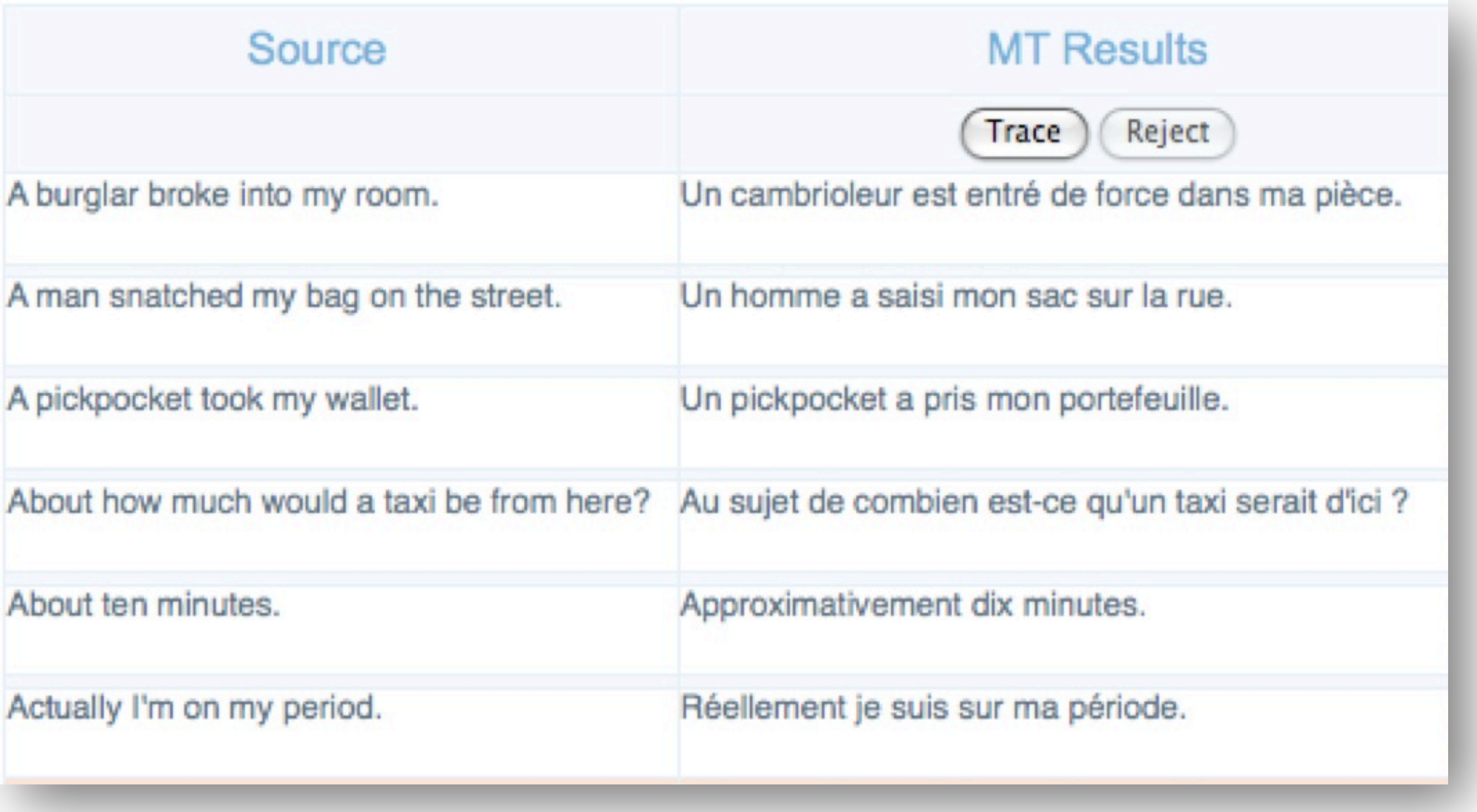
✓ TER Score: 0,00 (0,0/ 4,0)

- Source: actualy i’ m on my period .
- Best Ref: en fait j' ai mes règles .
- Orig Hyp: réellement je suis sur ma période .

```
REF:  en          fait j'   ai  mes règles  .
HYP:  réellement je   suis sur ma  période .
EVAL: S          S    S    S   S   S
SHFT:
```

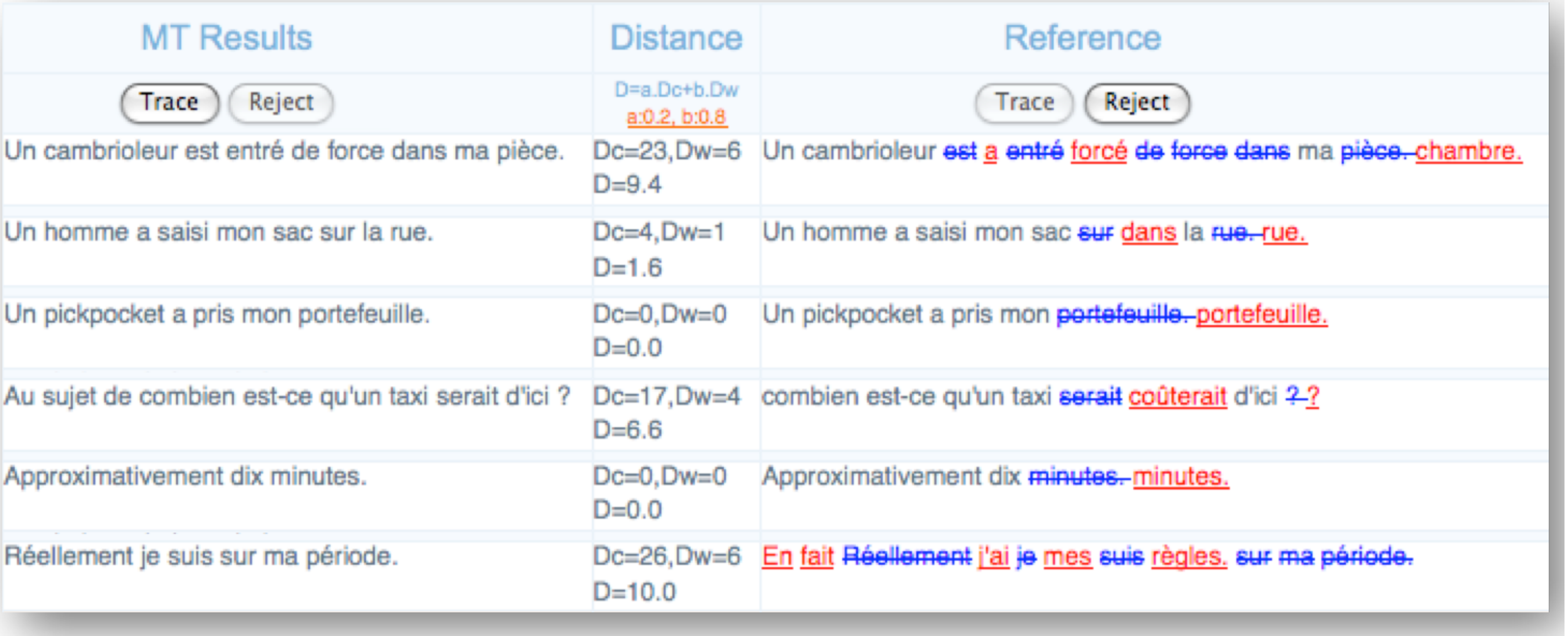
✓ TER Score: 85,71 (6,0/ 7,0)

TER with Sectra_w



Source	MT Results
A burglar broke into my room.	Un cambrioleur est entré de force dans ma pièce.
A man snatched my bag on the street.	Un homme a saisi mon sac sur la rue.
A pickpocket took my wallet.	Un pickpocket a pris mon portefeuille.
About how much would a taxi be from here?	Au sujet de combien est-ce qu'un taxi serait d'ici ?
About ten minutes.	Approximativement dix minutes.
Actually I'm on my period.	Réellement je suis sur ma période.

TER with Sectra_w



MT Results	Distance	Reference
Trace Reject	D=a.Dc+b.Dw a:0.2, b:0.8	Trace Reject

Un cambrioleur est entré de force dans ma pièce.	Dc=23,Dw=6
	D=9.4
		Un cambrioleur ~~est~~ a ~~entré~~ forcé ~~de force dans~~ ma ~~pièce.~~ chambre.

Un homme a saisi mon sac sur la rue.	Dc=4,Dw=1
	D=1.6
		Un homme a saisi mon sac ~~sur~~ dans la ~~rue.~~ rue.

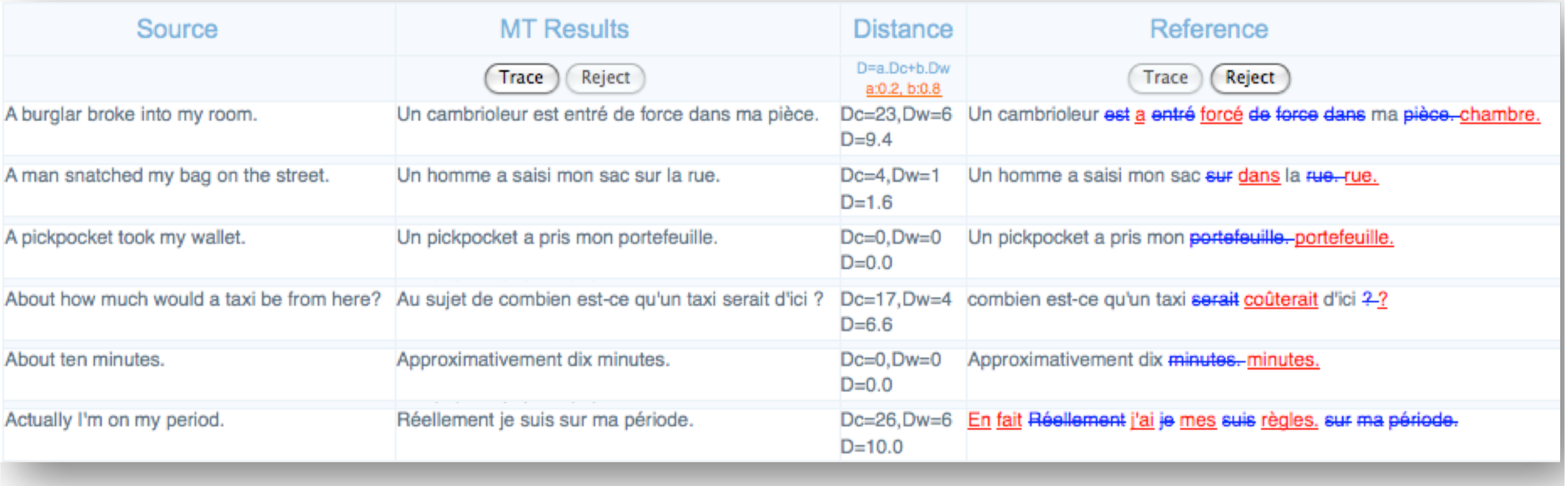
Un pickpocket a pris mon portefeuille.	Dc=0,Dw=0
	D=0.0
		Un pickpocket a pris mon ~~portefeuille.~~ portefeuille.

Au sujet de combien est-ce qu'un taxi serait d'ici ?	Dc=17,Dw=4
	D=6.6
		combien est-ce qu'un taxi ~~serait~~ coûterait d'ici ~~?~~ ?

Approximativement dix minutes.	Dc=0,Dw=0
	D=0.0
		Approximativement dix ~~minutes.~~ minutes.

Réellement je suis sur ma période.	Dc=26,Dw=6
	D=10.0
		En fait ~~Réellement~~ j'ai ~~je~~ mes ~~suis~~ règles. ~~sur ma période.~~

TER with Sectra_w



Source	MT Results	Distance	Reference
	Trace Reject	D=a.Dc+b.Dw a:0.2, b:0.8	Trace Reject

A burglar broke into my room.	Un cambrioleur est entré de force dans ma pièce.	Dc=23,Dw=6
		D=9.4
			Un cambrioleur ~~est~~ a ~~entré~~ forcé ~~de~~ ~~force~~ ~~dans~~ ma ~~pièce.~~ chambre.

A man snatched my bag on the street.	Un homme a saisi mon sac sur la rue.	Dc=4,Dw=1
		D=1.6
			Un homme a saisi mon sac ~~sur~~ dans la ~~rue.~~ rue.

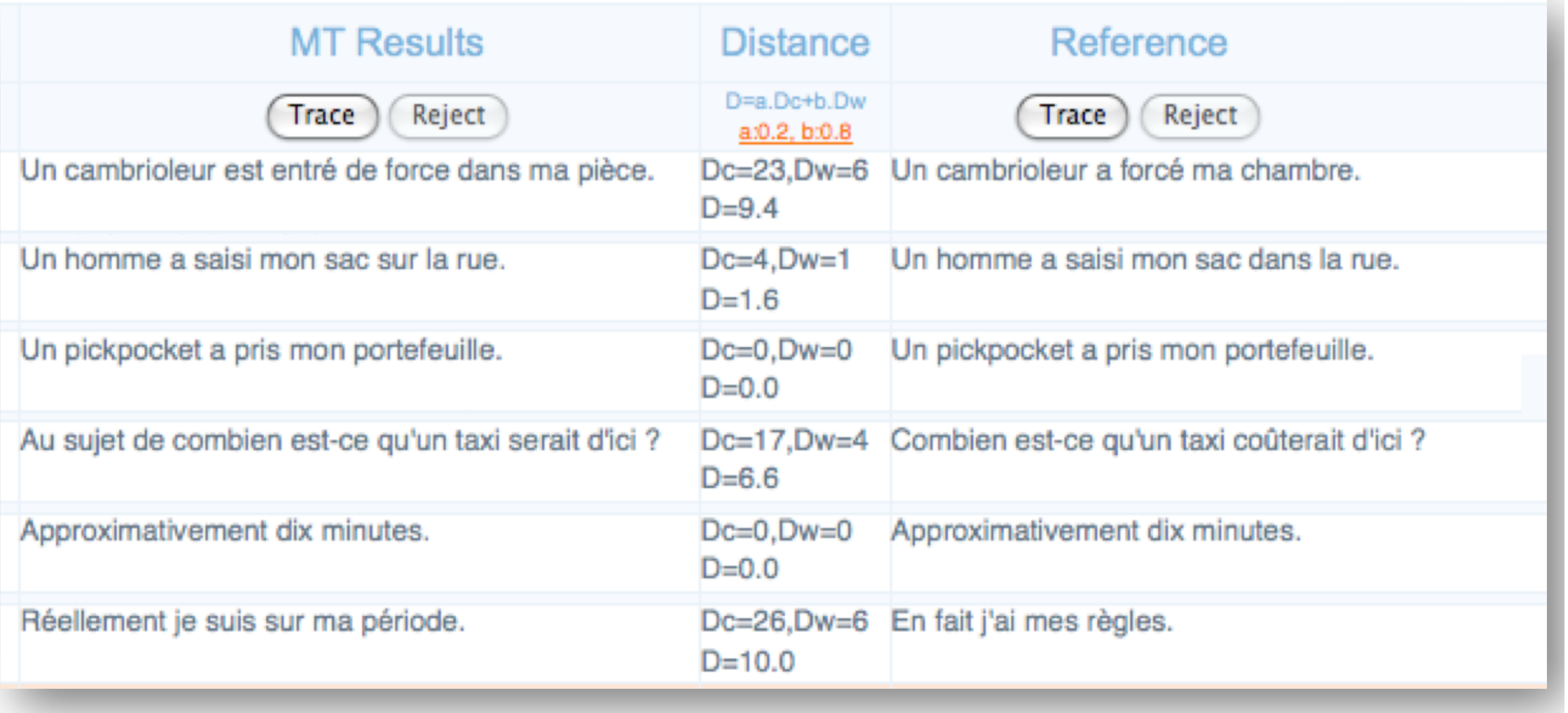
A pickpocket took my wallet.	Un pickpocket a pris mon portefeuille.	Dc=0,Dw=0
		D=0.0
			Un pickpocket a pris mon ~~portefeuille.~~ portefeuille.

About how much would a taxi be from here?	Au sujet de combien est-ce qu'un taxi serait d'ici ?	Dc=17,Dw=4
		D=6.6
			combien est-ce qu'un taxi ~~serait~~ coûterait d'ici ~~?~~ ?

About ten minutes.	Approximativement dix minutes.	Dc=0,Dw=0
		D=0.0
			Approximativement dix ~~minutes.~~ minutes.

Actually I'm on my period.	Réellement je suis sur ma période.	Dc=26,Dw=6
		D=10.0
			En fait ~~Réellement~~ j'ai ~~je~~ mes ~~suis~~ règles. ~~sur~~ ~~ma~~ ~~période.~~

TER with Sectra_w



MT Results	Distance	Reference
Trace Reject	D=a.Dc+b.Dw a:0.2, b:0.8	Trace Reject

Un cambrioleur est entré de force dans ma pièce.	Dc=23,Dw=6
	D=9.4
		Un cambrioleur a forcé ma chambre.

Un homme a saisi mon sac sur la rue.	Dc=4,Dw=1
	D=1.6
		Un homme a saisi mon sac dans la rue.

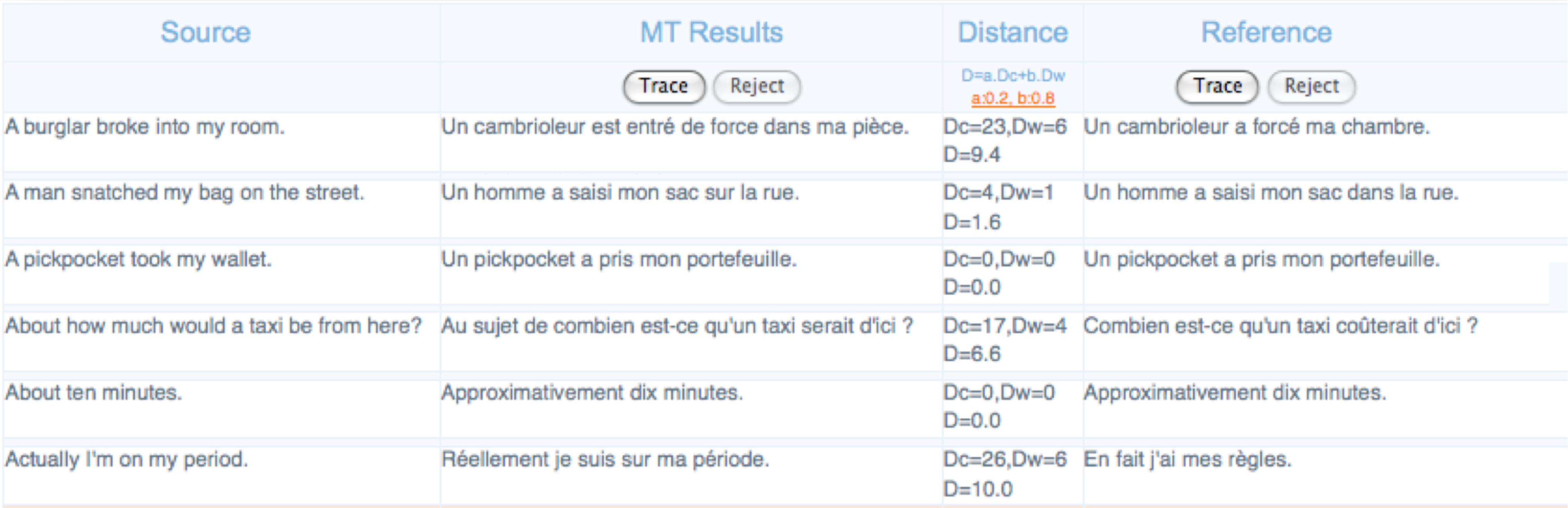
Un pickpocket a pris mon portefeuille.	Dc=0,Dw=0
	D=0.0
		Un pickpocket a pris mon portefeuille.

Au sujet de combien est-ce qu'un taxi serait d'ici ?	Dc=17,Dw=4
	D=6.6
		Combien est-ce qu'un taxi coûterait d'ici ?

Approximativement dix minutes.	Dc=0,Dw=0
	D=0.0
		Approximativement dix minutes.

Réellement je suis sur ma période.	Dc=26,Dw=6
	D=10.0
		En fait j'ai mes règles.

TER with Sectra_w



Source	MT Results	Distance	Reference
		D=a.Dc+b.Dw a:0.2, b:0.8	

A burglar broke into my room.	Un cambrioleur est entré de force dans ma pièce.	Dc=23,Dw=6
		D=9.4
			Un cambrioleur a forcé ma chambre.

A man snatched my bag on the street.	Un homme a saisi mon sac sur la rue.	Dc=4,Dw=1
		D=1.6
			Un homme a saisi mon sac dans la rue.

A pickpocket took my wallet.	Un pickpocket a pris mon portefeuille.	Dc=0,Dw=0
		D=0.0
			Un pickpocket a pris mon portefeuille.

About how much would a taxi be from here?	Au sujet de combien est-ce qu'un taxi serait d'ici ?	Dc=17,Dw=4
		D=6.6
			Combien est-ce qu'un taxi coûterait d'ici ?

About ten minutes.	Approximativement dix minutes.	Dc=0,Dw=0
		D=0.0
			Approximativement dix minutes.

Actually I'm on my period.	Réellement je suis sur ma période.	Dc=26,Dw=6
		D=10.0
			En fait j'ai mes règles.

Precision & Recall

- Precision
  - fraction of retrieved instances that are relevant

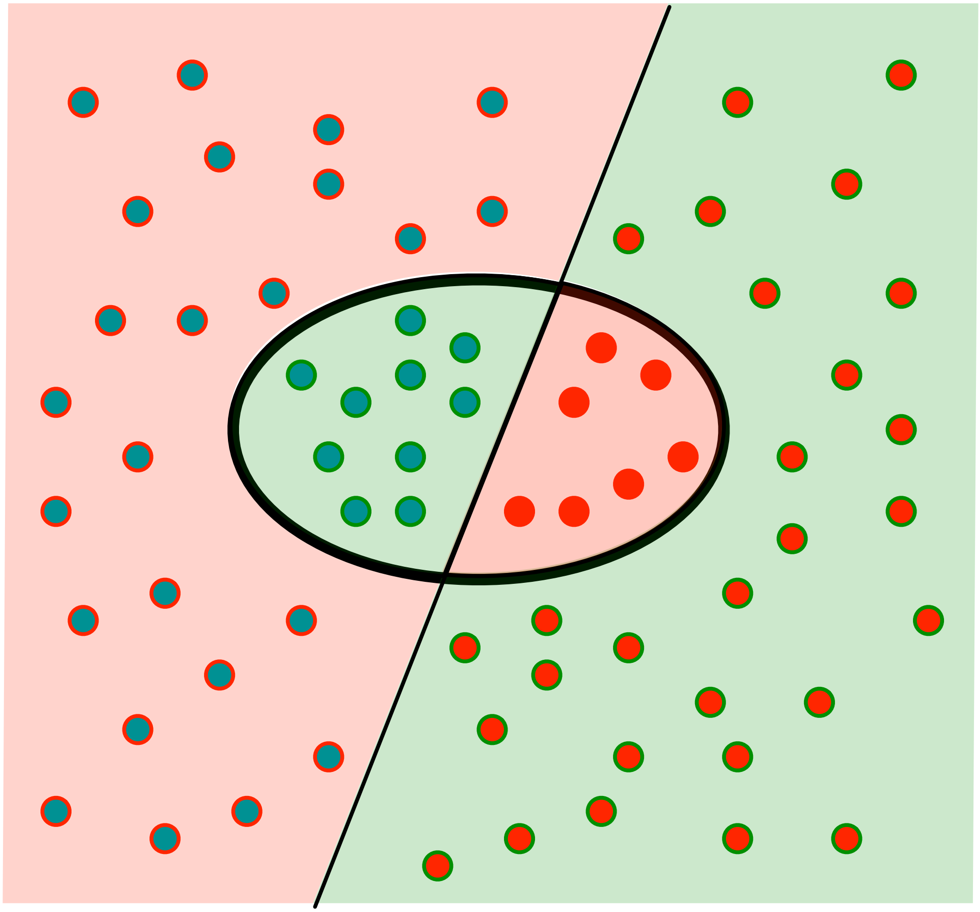
$$P = \frac{\text{\# of relevant answers}}{\text{\# of answers}}$$

- Recall
  - fraction of relevant instances that are retrieved

$$R = \frac{\text{\# of relevant answers}}{\text{\# of relevant instances}}$$

- *Example*

$$P = \frac{10}{17} = 0.58 \quad R = \frac{10}{34} = 0.29$$



● ● relevant instances *(33)* ● ● irrelevant instances

● ● answers *(16)*

● relevant answers *(9)* ● irrelevant answers false positives

● false negatives

Precision-oriented measures

Proportion of lexical units (n-grams) in the hypothesis covered by the reference(s) translation

- **BLEU** (Bilingual Evaluation Understudy) [Papinieni et al., 2001]
  - Modified precision (1 to 4 grams), geometric mean, brevity penalty
- **NIST** [Doddington, 2002]
  - N-gram informativeness (1 to 5 grams), arithmetic mean, brevity penalty
- **WNM** [Babych & Hartley, 2004]
  - Variant of BLEU which weights n-grams according to their statistical salience estimated out from a large monolingual corpus

BLEU: modified n-gram precision

- Definition
  - Count the number of occurrences of each candidate n-gram in the hypothesis and count their maximum number of occurrences in the associated reference(s)
  - Clip the candidate n-gram counts by their maximum number in the associated reference(s)
  - Sum the clipped count for all n-grams and divide by the total number of candidate n-grams

$$P_n = \frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} Count_{clip}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} Count(n-gram')}$$

BLEU: modified n-gram precision

- Example 1 on unigrams
  - Hypothesis
    - it is a guide to action which ensures that the military always obeys the commands of the party .
  - References
    - it is a guide to action that ensures that the military will forever heed party commands . (2 “that”)
    - it is the guiding principle which guarantees the military forces always being under the command of the party . (4 “the”)
    - it is the practical guide for the army always to heed the directions of the party . (3 “the”)

BLEU: modified n-gram precision

- Example 1 on unigrams (cont.)

Candidate words	Count	Max_ref_count	*Count*$_{clip}$
it	1	1	1
is	1	1	1
a	1	1	1
guide	1	1	1
to	1	1	1
action	1	1	1
which	1	1	1
ensure	1	1	1
that	1	2	1
military	1	1	1
always	1	1	1
obeys	1	0	0
the	3	4	3
commands	1	1	1
of	1	1	1
party	1	1	1
sum	18	/	17

$$P_1 = \frac{17}{18}$$

BLEU: modified n-gram precision

- Example 2 on unigrams
  - Hypothesis
    - it is to insure the troops forever hearing the activity guidebook that party direct .
- References
    - it is a guide to action that ensures that the military will forever heed party commands . (2 “that”)
    - it is the guiding principle which guarantees the military forces always being under the command of the party . (4 “the”)
    - it is the practical guide for the army always to heed the directions of the party . (2 “the”)

BLEU: modified n-gram precision

- Example 2 on unigrams (cont.)

Candidate words	Count	Max_ref_count	$Count_{clip}$
it	1	1	1
is	1	1	1
to	1	1	1
insure	1	0	0
the	2	4	2
troops	1	0	0
forever	1	1	1
hearing	1	0	0
activity	1	0	0
guidebook	1	0	0
that	1	2	1
party	1	1	1
direct	1	0	0
sum	14	/	8

$$P_1 = \frac{8}{14}$$

BLEU: hypotheses brevity penalty

- definition
  - Hypothesis longer than references already penalized with modified precision (Countclip/Count)
  - Need to penalize shorter hypotheses

No penalty when the hypothesis length is the same as any reference

$$r = \sum_{\mathrm{C} \in \{candidats\}} \text{best reference match for C}$$

- let r be the test corpus' effective reference length

$$c = \sum_{\mathrm{C} \in \{candidats\}} \text{length of C}$$

- let c be the total length of the hypothesis corpus
- Brevity Penalty

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{(1-{}^{r}/_{c})}, & \text{if } c \leq r \end{cases}$$

BLEU: the formula

- BLEU is computed as follows:

$$\mathrm{BLEU} = BP \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

  - where
    - $N = 4$ and $w_n = 1/N$
  - $\mathrm{BLEU} \in [0..1]$

BLEU: example

- Reference: the green house was right in front of the lake .
- Translation 0: the green house was right in front of the lake .

```
For N-Gram (green ): 1
For N-Gram (house ): 1
For N-Gram (was ): 1
For N-Gram (right ): 1
For N-Gram (in ): 1
For N-Gram (front ): 1
For N-Gram (of ): 1
For N-Gram (the ): 2
For N-Gram (lake ): 1
```

```
For N-Gram (the green ): 1
For N-Gram (green house ): 1
For N-Gram (house was ): 1
For N-Gram (was right ): 1
For N-Gram (right in ): 1
For N-Gram (in front ): 1
For N-Gram (front of ): 1
For N-Gram (of the ): 1
For N-Gram (the lake ): 1
```

```
For N-Gram (the green house ): 1
For N-Gram (green house was ): 1
For N-Gram (house was right ): 1
For N-Gram (was right in ): 1
For N-Gram (right in front ): 1
For N-Gram (in front of ): 1
For N-Gram (front of the ): 1
For N-Gram (of the lake ): 1
```

```
For N-Gram (the green house was ): 1
For N-Gram (green house was right ): 1
For N-Gram (house was right in ): 1
For N-Gram (was right in front ): 1
For N-Gram (right in front of ): 1
For N-Gram (in front of the ): 1
For N-Gram (front of the lake ): 1
```

```
Precision 1-gram: 1.00 = 10/10
Precision 2-gram: 1.00 = 9/9
Precision 3-gram: 1.00 = 8/8
Precision 4-gram: 1.0 = 7/7
Weighted Precision: 1.00
Brevity Penalty: 1.00
-------------------------
BLEU = 1.00
```

MyBleuHB & mteval-v11b.pl

Back to subjective evaluation

- Fluency evaluation for the 3 following translations

	Fluency
a green house was by the lake shore .	5
the green house was by the lake shore .	5
the green potato right in front of the lake was right .	3~1

Score	Fluency
5	Flawless English
4	Good
3	Non-Native
2	Disfluent
1	Incomprehensible

Back to subjective evaluation

- Adequacy evaluation given reference
  - the green house was right in front of the lake .

	Adequacy
a green house was by the lake shore .	5~4
the green house was by the lake shore .	5
the green potato right in front of the lake was right .	1

Score	Adequacy
5	All information
4	Most
3	Much
2	Little
1	None

BLEU: example

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .

```
For N-Gram (a ): 0
For N-Gram (green ): 1
For N-Gram (house ): 1
For N-Gram (was ): 1
For N-Gram (by ): 0
For N-Gram (the ): 1
For N-Gram (lake ): 1
For N-Gram (shore ): 0
```

```
For N-Gram (a green ): 0
For N-Gram (green house ): 1
For N-Gram (house was ): 1
For N-Gram (was by ): 0
For N-Gram (by the ): 0
For N-Gram (the lake ): 1
For N-Gram (lake shore ): 0
```

```
For N-Gram (a green house ): 0
For N-Gram (green house was ): 1
For N-Gram (house was by ): 0
For N-Gram (was by the ): 0
For N-Gram (by the lake ): 0
For N-Gram (the lake shore ): 0
```

```
For N-Gram (a green house was ): 0
For N-Gram (green house was by ): 0
For N-Gram (house was by the ): 0
For N-Gram (was by the lake ): 0
For N-Gram (by the lake shore ): 0
```

```
Precision 1-gram: 0.625000 = 5/8
Precision 2-gram: 0.428571 = 3/7
Precision 3-gram: 0.166667 = 1/6
Precision 4-gram: 0.000000 = 0/5
Weighted Precision: 0.000000
   (because 4-gram precision = 0)
Brevity Penalty: 0.778801
-------------------------
BLEU = 0.000000
```

BLEU: example

- Reference: the green house was right in front of the lake .
- Translation 2: the green house was by the lake shore .

```
For N-Gram (green ): 1
For N-Gram (house ): 1
For N-Gram (was ): 1
For N-Gram (by ): 0
For N-Gram (the ): 2
For N-Gram (lake ): 1
For N-Gram (shore ): 0
```

```
For N-Gram (the green ): 1
For N-Gram (green house ): 1
For N-Gram (house was ): 1
For N-Gram (was by ): 0
For N-Gram (by the ): 0
For N-Gram (the lake ): 1
For N-Gram (lake shore ): 0
```

```
For N-Gram (the green house ): 1
For N-Gram (green house was ): 1
For N-Gram (house was by ): 0
For N-Gram (was by the ): 0
For N-Gram (by the lake ): 0
For N-Gram (the lake shore ): 0
```

```
For N-Gram (the green house was ): 1
For N-Gram (green house was by ): 0
For N-Gram (house was by the ): 0
For N-Gram (was by the lake ): 0
For N-Gram (by the lake shore ): 0
```

```
Precision 1-gram: 0.750000 = 6/8
Precision 2-gram: 0.571429 = 4/7
Precision 3-gram: 0.333333 = 2/6
Precision 4-gram: 0.200000 = 1/5
Weighted Precision: 0.411134
Brevity Penalty: 0.778801
-------------------------
BLEU = 0.320191
```

BLEU: example

- Reference: the green house was right in front of the lake .
- Trans. 3: the green potato right in front of the lake was right .

```
For N-Gram (green ): 1
For N-Gram (potato ): 0
For N-Gram (in ): 1
For N-Gram (front ): 1
For N-Gram (of ): 1
For N-Gram (the ): 2
For N-Gram (lake ): 1
For N-Gram (was ): 1
For N-Gram (right ): 1
```

```
For N-Gram (the green ): 1
For N-Gram (green potato ): 0
For N-Gram (potato right ): 0
For N-Gram (right in ): 1
For N-Gram (in front ): 1
For N-Gram (front of ): 1
For N-Gram (of the ): 1
For N-Gram (the lake ): 1
For N-Gram (lake was ): 0
For N-Gram (was right ): 1
```

```
For N-Gram (the green potato ): 0
For N-Gram (green potato right ): 0
For N-Gram (potato right in ): 0
For N-Gram (right in front ): 1
For N-Gram (in front of ): 1
For N-Gram (front of the ): 1
For N-Gram (of the lake ): 1
For N-Gram (the lake was ): 0
For N-Gram (lake was right ): 0
```

```
For N-Gram (the green potato right ): 0
For N-Gram (green potato right in ): 0
For N-Gram (potato right in front ): 0
For N-Gram (right in front of ): 1
For N-Gram (in front of the ): 1
For N-Gram (front of the lake ): 1
For N-Gram (of the lake was ): 0
For N-Gram (the lake was right ): 0
```

```
Precision 1-gram: 0.818182 = 9/11
Precision 2-gram: 0.700000 = 7/10
Precision 3-gram: 0.444444 = 4/9
Precision 4-gram: 0.375000 = 3/8
Weighted Precision: 0.555839
Brevity Penalty: 1.000000
------------------------
BLEU = 0.555839
```

BLEU: example

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Translation 2: the green house was by the lake shore .
- Translation 3: the green potato right in front of the lake was right .

	WP	BP	BLEU
T1	0.000000	0.778801	**0.000000**
T2	0.411134	0.778801	**0.320191**
T3	0.555839	1.000000	**0.555839**

- Don’t we have a problem!!!!
  - T1 acceptable (one word changed compared to T2)
  - T3 wrong and nonsense

NIST: n-gram information weight

- Definition
  - With BLEU all n-grams are equally important
  - NIST associate an information weight to each n-gram of the reference set

$$Info(w_1 w_2 \dots w_n) = \log_2\left(\frac{\text{the \# of occurrences of } w_1 w_2 \dots w_{n-1}}{\text{the \# of occurrences of } w_1 w_2 \dots w_n}\right)$$

- for a unigram w_1:

  the # of occurrences = the # of occurrences in the reference

NIST: hypotheses brevity penalty

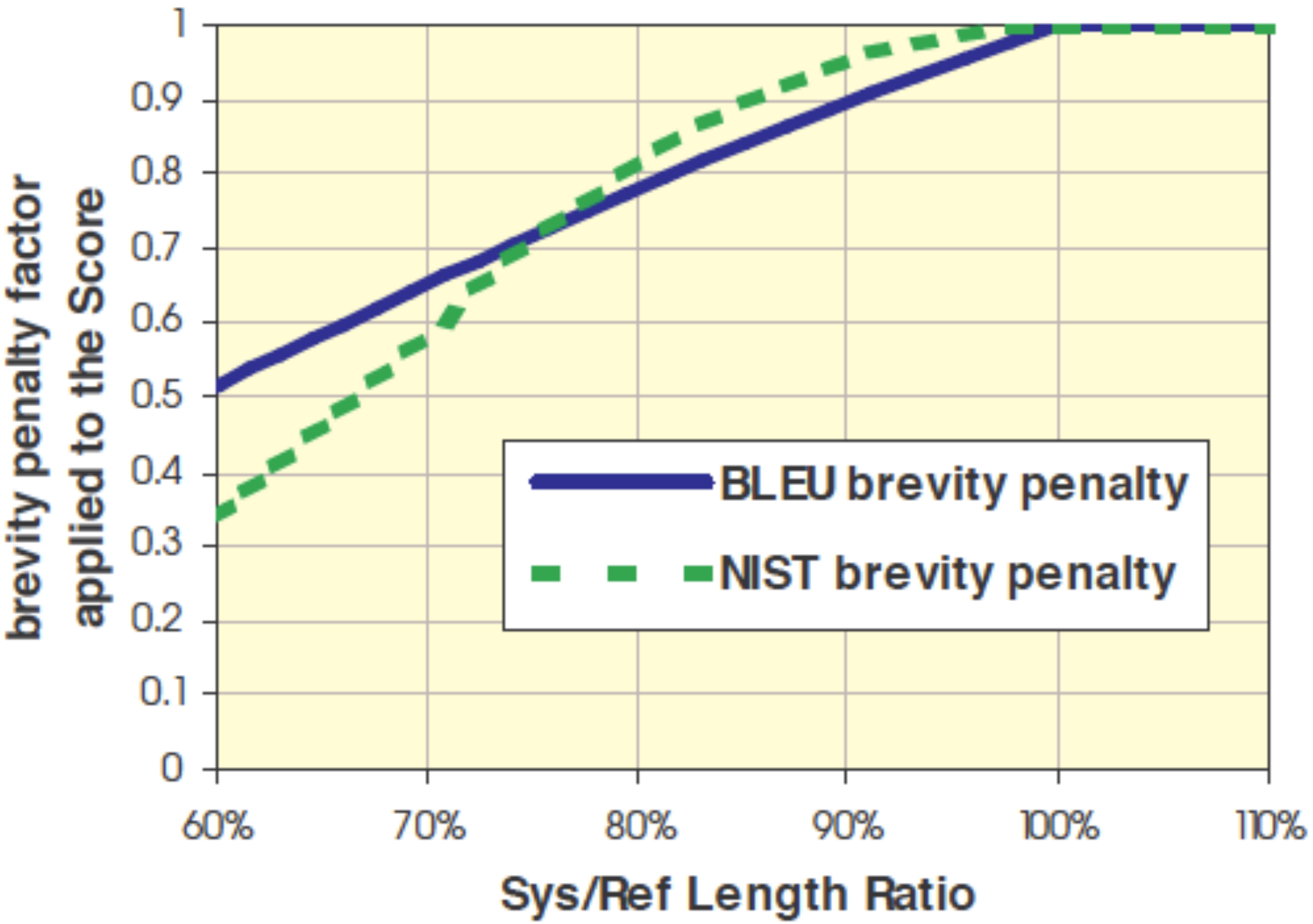
- Definition
  - New BP to minimize the impact on the score of small variations in the length of a translation
  - It reduces the contributions of length variations to the score for small variations

$$BP = \exp\left\{\beta \log^2\left[\min\left(\frac{L_{sys}}{\bar{L}_{ref}}, 1\right)\right]\right\}$$

  - where
    - β is chosen to make the brevity penalty factor $=\ 0.5$ when the # of words in the system output is $2/3$ of the average # of words in the reference translation
    - $\overline{L}_{ref}$ $=$ the average number of words in a reference translation, averaged over all reference translations
    - L_{sys} $=$ the number of words in the translation being scored

BLEU vs NIST: Brevity Penalty

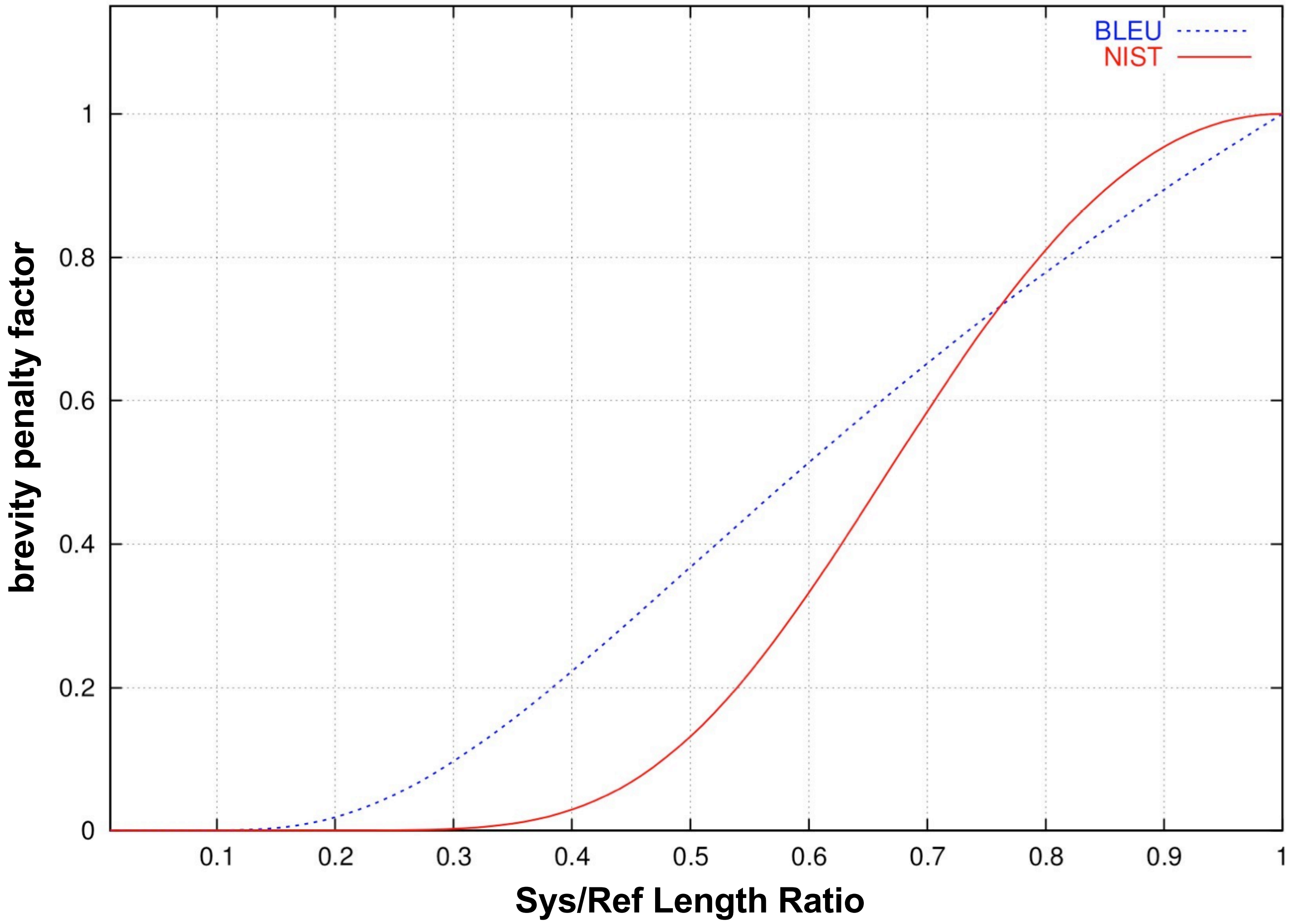
- Hypo/Ref Length Ratio ≈ 0.85



from [Doddington, 2002]

BLEU vs NIST: Brevity Penalty

- 0 < Hypo(Sys)/Ref Length Ratio ≤ 1



NIST: the formula

- NIST is computed as follows:

$$NIST = BP \cdot \sum_{n=1}^{N} \left\{ \frac{\sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{that co-occur}}} Info(w_1 \ldots w_n)}{\sum_{\substack{\text{all } w_1 \ldots w_n \\ \text{in hypothesis}}} (1)} \right\}$$

- Where
  - $N \;=\;$ 4 at least

- $NIST \;\in\; [0..+\infty[$ *(*$[0..15[$ *in practice)*

NIST: example

$$Info(w_1 w_2 \dots w_n) = \log_2\left(\frac{\text{the \# of occurrences of } w_1 w_2 \dots w_{n-1}}{\text{the \# of occurrences of } w_1 w_2 \dots w_n}\right)$$

- Reference: the green ~~house was~~ right in front of the lake . (11 1-grams)
- Translation 1: a green house was by the lake shore .
- Co-occurring n-grams
  - 1-grams: ‘the’, ‘green’, ‘house’, ‘was’, ‘lake’, ‘.’
  - 2-grams: ‘green house’, ‘house was’, ‘the lake’
    - the green ~~house was~~ right in front of the lake .
    - a green ~~house was~~ by the lake shore .
  - 3-gram: “green house was”
    - the green house was right in front of the lake .
    - a green house was by the lake shore .
- Info
  - Info(the)=$\log_2(11/2) = 2.4594$
  - Info(green)=Info(house)=Info(was)=Info(lake)=Info(.)= $\log_2(11/1) = 3.4594$
  - Info(green house)=Info(~~house was~~)= $\log_2(1/1) = 0.0000$
  - Info(the lake)= $\log_2(2/1) = 1.0000$
  - Info(green ~~house was~~)= $\log_2(1/1) = 0.0000$

NIST: Example

- $$\text{NIST 1-gram} = \frac{\text{Info(the)+Info(green)+Info(house)+Info(was)+Info(lake)+Info(.)}}{9_{\textit{(\# 1-grams in hypothesis)}}}$$

- $$\text{NIST 2-gram} = \frac{\text{Info(green house)+Info(house was)+Info(the lake)}}{8_{\textit{(\# 2-grams in hypothesis)}}}$$

- $$\text{NIST 2-gram} = \frac{\text{Info(green house was)}}{7_{\textit{(\# 3-grams in hypothesis)}}}$$

- Penalty
  - $\beta = 4.2162; ratio\ hypo/ref = 9/11 = 0.8181$
  - $BP = \exp(\beta \cdot \log^2(ratio)) = 0.8439$

NIST: example

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .

- `NIST score = 1.9579`
- `Brevity Penalty = 0.8439`

```
# ----------------------------------------------------
Individual N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  2.1951   0.1250   0.0000   0.0000   0.0000
# -----------------------------------------------------
Cumulative N-gram scoring × BP
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  1.8524   0.1055   0.0000   0.0000   0.0000  ∑=1.9579
```

mteval-v11b.pl

NIST: example

- Reference: the green house was right in front of the lake .
- Translation 2: the green house was by the lake shore .

- `NIST score = 2.2940`
- `Brevity Penalty = 0.8439`

```
# ----------------------------------------
Individual N-gram scoring with BP
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  2.0830   0.2110   0.0000   0.0000   0.0000
# ----------------------------------------
Individual N-gram scoring x BP
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  2.0830   2.2940   2.2940   2.2940   2.2940
```

NIST: example

- Reference: the green house was right in front of the lake .
- Trans. 3: the green potato right in front of the lake was right .

- `NIST score = 2.8980`
- `Brevity Penalty = 1.0000`

```
# -------------------------------------------
Individual N-gram scoring with BP
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  2.7162   0.1818   0.0000   0.0000   0.0000
# -------------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  2.7162   2.8980   2.8980   2.8980   2.8980
```

NIST: example

- Reference: the green house was right in front of the lake .
- Translation 1: a green house was by the lake shore .
- Translation 2: the green house was by the lake shore .
- Translation 3: the green potato right in front of the lake was right .

	BLEU
T1	1.9579
T2	2.2940
T3	2.8980

- Don’t we have a problem!!!!
  - T1 acceptable (one word changed compared to T2)
  - T3 wrong and nonsense

NIST: example

- Reference: the green house was right in front of the lake .
- Translation 0: the green house was right in front of the lake .

- `NIST score = 3.2776`
- `Brevity Penalty = 1.0000`

```
# ----------------------------------------
Individual N-gram scoring with BP
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  3.2776   0.2000   0.0000   0.0000   0.0000
# ----------------------------------------
Cumulative N-gram scoring
        1-gram   2-gram   3-gram   4-gram   5-gram
        ------   ------   ------   ------   ------
 NIST:  3.2776   3.4776   3.4776   3.4776   3.4776
```

Recall-oriented measures

Proportion of the lexical unit in the reference translation(s) covered by the hypothesis

- **ROUGE** (Recall-Oriented Understudy for Gisiting Evaluation) [Lin & Och, 2004]
  - Lexical recall among n-grams (1 to 4 grams); allows for stemming and discontinuous matching (skip n-grams)
- **CDER** (Cover/Disjoint Error Rate) [Leusch et al., 2006]
  - Recall oriented measure modeling block reordering; movements of word blocks as an edit operation

Measures balancing precision & recall

- Precision & recall combination

F_1 score $F_1 = 2 \cdot \frac{P \cdot R}{P + R}$, F_β score $F_\beta = (1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R}$

- **GTM** (General Text Matcher) [Melamed et al., 2003; Turian et al., 2003]
  - F-measure; adjusted importance of n-grams matching
- **METEOR** [Banerjee & Lavie, 2005]
  - F-measure based on 1-gram alignment & word ordering;
    + stemming & synonymy through WordNet
- **BLANC** [Lita et al., 2005]
  - Family of trainable n-gram based metrics; variable size non-continuous word sequences
- **SIA** (Stochastic Iterative Alignment) [Liu & Gileda, 2006]
  - Loose sequence alignment enhanced with alignment scores, stochastic word matching and iterative alignment scheme

CLIPS at IWSLT 2004

- Setting
  - Language pair
    - Japanese ⟶ English
  - Domain
    - Tourism
  - Systems
    - Systran Web & Systran Professionnal Premium (PP) V5
      - used at that time as baseline systems

clips-1	Systran Web V5
clips-2	Systran PP V5 with original dictionaries
clips-3	Systran PP V5 with original and user dictionaries

CLIPS at IWSLT 2004

Results

- Subjective evaluation of **clips-3**
  - non-native English > Fluency > disfluent English
  - much > Adequacy > little
- Objective evaluation (score$_{rank}$)

	BLEU	GMT	NIST	PER	WER
clips-3	0.1320_1	0.5687_1	5.6476_1	0.5978_1	0.7304_1
clips-2	0.1311_2	0.5672_2	5.6096_2	0.6012_2	0.7349_2
clips-1	0.0810_3	0.5116_3	4.1935_3	0.7179_3	0.8726_3

  - systems ranked as expected!

CLIPS at IWSLT 2004

- Errors
  - Bad translation when subject is omitted
    - ここ で 降り ます。 → **It gets** off here.
      - (koko de orimasu) I will get off here.
  - Euphemistic utterance が translated by “but”
    - 両替 を し たい の です が。 → It is to like to exchange **but**.
      - (ryoukake o shitai no desu ga) I would like to change money.
  - Question word order
    - 入場 料 は いくら です か。 → Is admission fee **how much**?
      - (nyuujouryou wa ikura desu ka) How much is the admission fee?
  - Requests or invitations
    - 一緒 に 行き ましょ う。 → **It will** go together.
      - (isshoni ikimashou) Let’s go together.

CLIPS at IWSLT 2004

TER

- Best Ref: i will get off here .
- Orig Hyp: it gets off here .

```
REF:  i will get  off here .
HYP:  * it   gets off here .
EVAL: D S    S
SHFT:
TER Score:  50,00 (  3,0/  6,0)
```

CLIPS at IWSLT 2004

- TER
  - Best Ref: i would like to exchange money .
  - Orig Hyp: it is to like to exchange but .

```
REF:  ** i  would like to exchange money .
HYP:  it is to    like to exchange but   .
EVAL: I  S  S                      S
SHFT:
TER Score:  57,14 (  4,0/  7,0)
```

CLIPS at IWSLT 2004

- TER
  - Best Ref: how much is the admission fee ?
  - Orig Hyp: is admission fee how much ?

```
REF:    how much   is the admission fee   ?
HYP:  [ how much ] is *** admission fee @ ?
EVAL:                 D
SHFT: 1          1                      1
TER Score:  28,57 (  2,0/  7,0)
```

```
Shift [how, much] 3 words left
 REF:    how much   is the admission fee   ?
 HYP:  [ how much ] is *** admission fee @ ?
```

CLIPS at IWSLT 2004

- TER
  - Best Ref: let 's go together .
  - Orig Hyp: it will go together .

```
REF:  let 's   go together .
HYP:  it  will go together .
EVAL: S   S
SHFT:
TER Score:  40,00 (  2,0/  5,0)
```

CLIPS at IWSLT 2004

- Competitive objective evaluation
  - Systran PP V5 is fourth

	BLEU	GMT	NIST	PER	WER
JE_1	0.6306$_1$	0.6306$_2$	10.7201$_2$	0.2333$_1$	0.2631$_1$
JE_3	0.6190$_2$	0.8243$_1$	11.2541$_1$	0.2492$_1$	0.3056$_1$
JE_4	0.3970$_3$	0.6722$_3$	7.8893$_3$	0.4202$_3$	0.4857$_3$
CLIPS-3	0.1320$_4$	0.5687$_4$	5.6476$_4$	0.5978 $_4$	0.7304$_4$

CLIPS at IWSLT 2004

- Competitive objective evaluation with post-edited (PE) Systran outputs
  - A 5 score at subjective evaluation for both fluidity and adequacy

	BLEU	GMT	NIST	PER	WER
JE_1	0.6306$_1$	0.6306$_2$	10.7201$_2$	0.2333$_1$	0.2631$_1$
JE_3	0.6190$_2$	0.8243$_1$	11.2541$_1$	0.2492$_1$	0.3056$_1$
PE-CLIPS-3	0.4691$_-$	0.7777$_-$	9.9189$_-$	0.3236$_-$	0.3711$_-$
JE_4	0.3970$_3$	0.6722$_3$	7.8893$_3$	0.4202$_3$	0.4857$_3$
CLIPS-3	0.1320$_4$	0.5687$_4$	5.6476$_4$	0.5978 $_4$	0.7304$_4$

  - Scores improve but not enough post-eds still far from refs
    - system JE_4 beaten
    - PE-CLIPS-3 still to far from references to beat JE_1 & JE_3

Zied Elloumi, Hervé Blanchon

Gilles Sérasset & Laurent Besacier

MT SUMMIT 2015

METEOR FOR MULTIPLE TARGET LANGUAGES USING DBNARY

Outline

- **Situation**
  - METEOR, WordNet, **Dbnary**
- **“Dbnary Synsets” Extraction**
- **METEOR Scores on English**
  - WordNet *vs* Dbnary Synsets
- **Correlation with human judgment**
  - METEOR without Synset *vs* METEOR with “Dbnary Synsets”
- **Conclusion and Perspectives**

SITUATION

METEOR

- Introduced by *(Banerjee and Lavie, 2005*)
  - to overcome several weaknesses of BLEU *(Papineni, 2002)* and NIST *(Doddington, 2002)*
  - to better correlate with human judgment
- A 3-leveled mapping approach
  between a MT Hypothesis and one or several References
  - **surface forms overlap** of words
  - **stems** (lemma) **overlap** of surface forms
    - tool: a stemmer (lemmatizer) for the language
  - **synonymy overlap** through shared **WordNet Synsets**
    - resource: a WordNet for the language

METEOR Recent Extensions

- METEOR-NEXT *(Denkowski and Lavie, 2010a)*
  - to better correlate with HTER *(Snover & al., 2006)*
  - a 4th mapping level to accommodate **multi-word matches**
    - <u>resource</u>: a paraphrase database for the language
- METEOR Universal *(Denkowski and Lavie 2014)*
  - <u>tool</u>: automatic extraction of paraphrase tables and function word lists from bitexts
  - <u>resources</u>: paraphrase tables for English, Arabic, Czech, French, German, Spanish
  - parameter set (learned from human judgments)
- METEOR-WSD *(Apidianaki and Marie, 2015)*
  - to filter synonyms/paraphrases according to word senses
  - English references further disambiguated and annotated using Babelfly *(Moro et al., 2014)*

WordNet

- A large lexical database for English *(Fellbaum, 1998)*
- WordNet links **nouns**, **verbs**, **adjectives** and **adverbs** to sets of cognitive synonyms (Synsets)
- Different versions of WordNet in other languages (Arabic, French, …)
  - pro: important and a very useful resources
  - cons: not free and/or not available for every language

METEOR & WordNet

- Pro
  - synonym match increases the chance of the MT output words to match the reference words
- Cons
  - synonym match available only for English

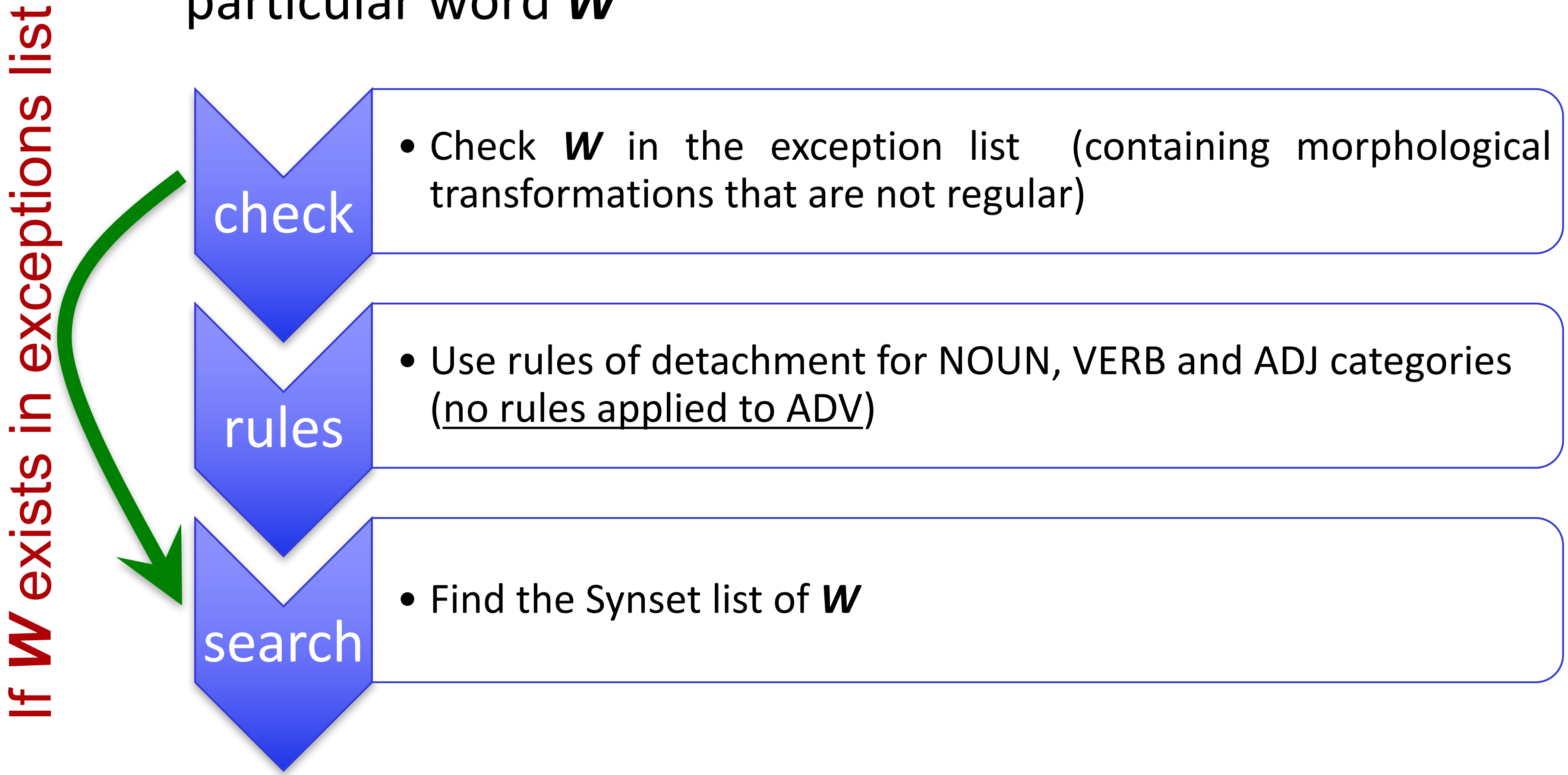
- Latest version of WordNet 3.0 = 117 659 Synsets

Categories	# of Synsets
Verb	13 767
Noun	82 115
Adverb	3 621
Adjective	18 156

Table 1. Number of Synsets in WordNet

METEOR & WordNet

- METEOR uses the **Morphy-7WN** function from WordNet to lemmatize forms
- **Morphy-7WN** uses a two-step process to find lemma of a particular word ***W***



- Check ***W*** in the exception list (containing morphological transformations that are not regular)

- Use rules of detachment for NOUN, VERB and ADJ categories (no rules applied to ADV)

- Find the Synset list of ***W***

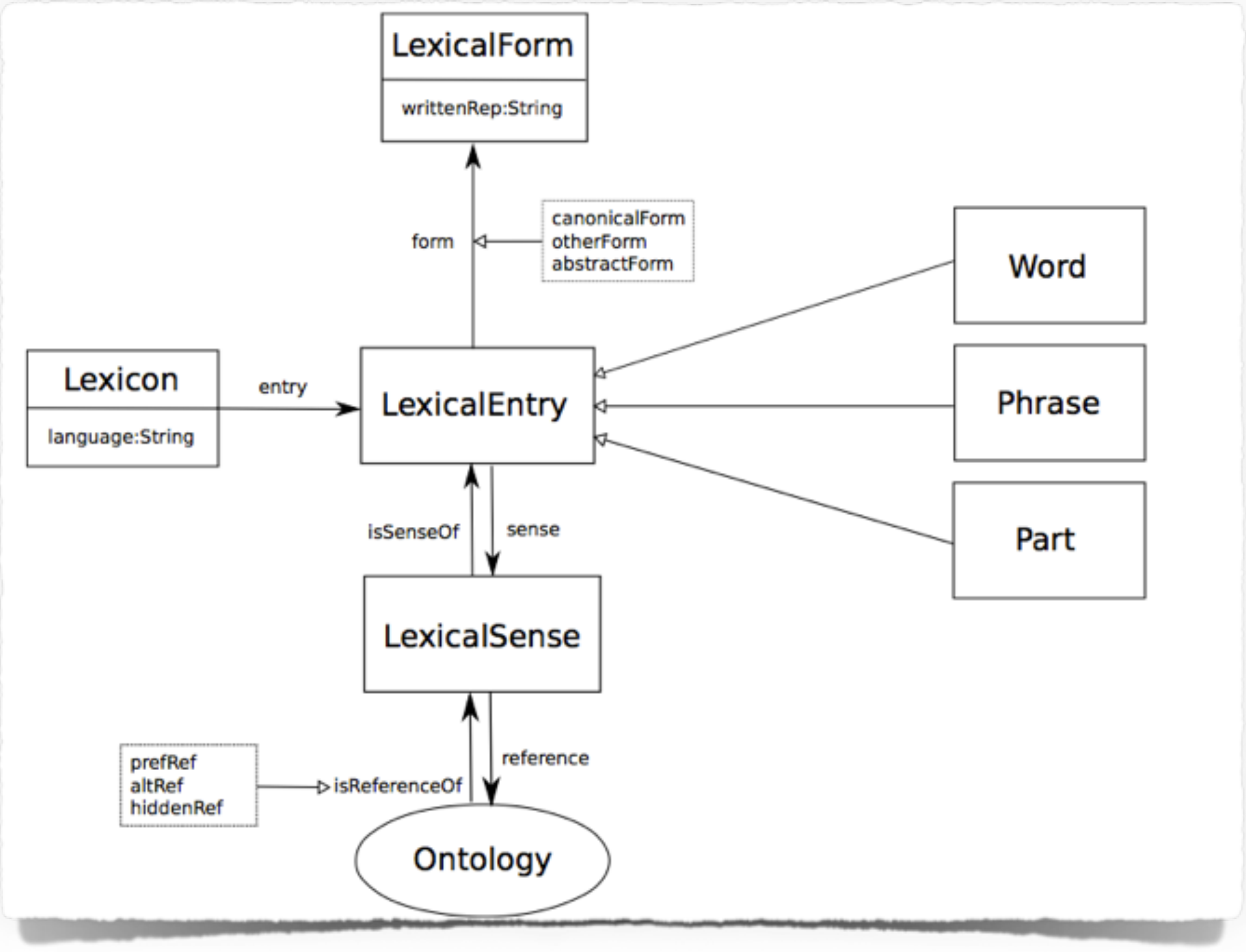
Dbnary (http://kaiko.getalp.org/about-dbnary/)

- What is it?
  - a multilingual lexical resource in RDF *(Klyne & Carroll, 2004)* collected at the LIG *(Sérasset, 2015)* and extracted from Wiktionary (currently 21 languages editions)
  - the lexical data is made available as LLOD (Linguistic Linked Open Data)
  - the lexicon structure is defined using the LEMON vocabulary *(McCrae et al., 2011)*
- Availability
  - downloadable files
    - queried locally using SPARQL
  - Linked Open Data directly accessible to browsers or applications
    - queried online using SPARQL

Wiktionary	the dictionary counterpart of Wikipedia
LEMON	a model for modeling lexicon and machine-readable dictionaries linked to the Semantic Web and the Linked Data cloud
SPARQL	a standard language for querying linked data

Dbnary: the dataset

- Core data
  - **Lexical Entries, Lexical Senses** and **Translations**
- Additional data
  - Semantically enriched Relations
    - **Translations**: attached to their source Lexical Sense when possible
    - **Lexico-semantic relations**: also attached to their source Lexical Sense
      - **syno**/anto-**nymy**, hypo/hyper-nymy
      - mero/holo-nymy, tropo-nymy
  - Morphology
    - Extensive representation of morphology (a set of "lemon:otherForm")

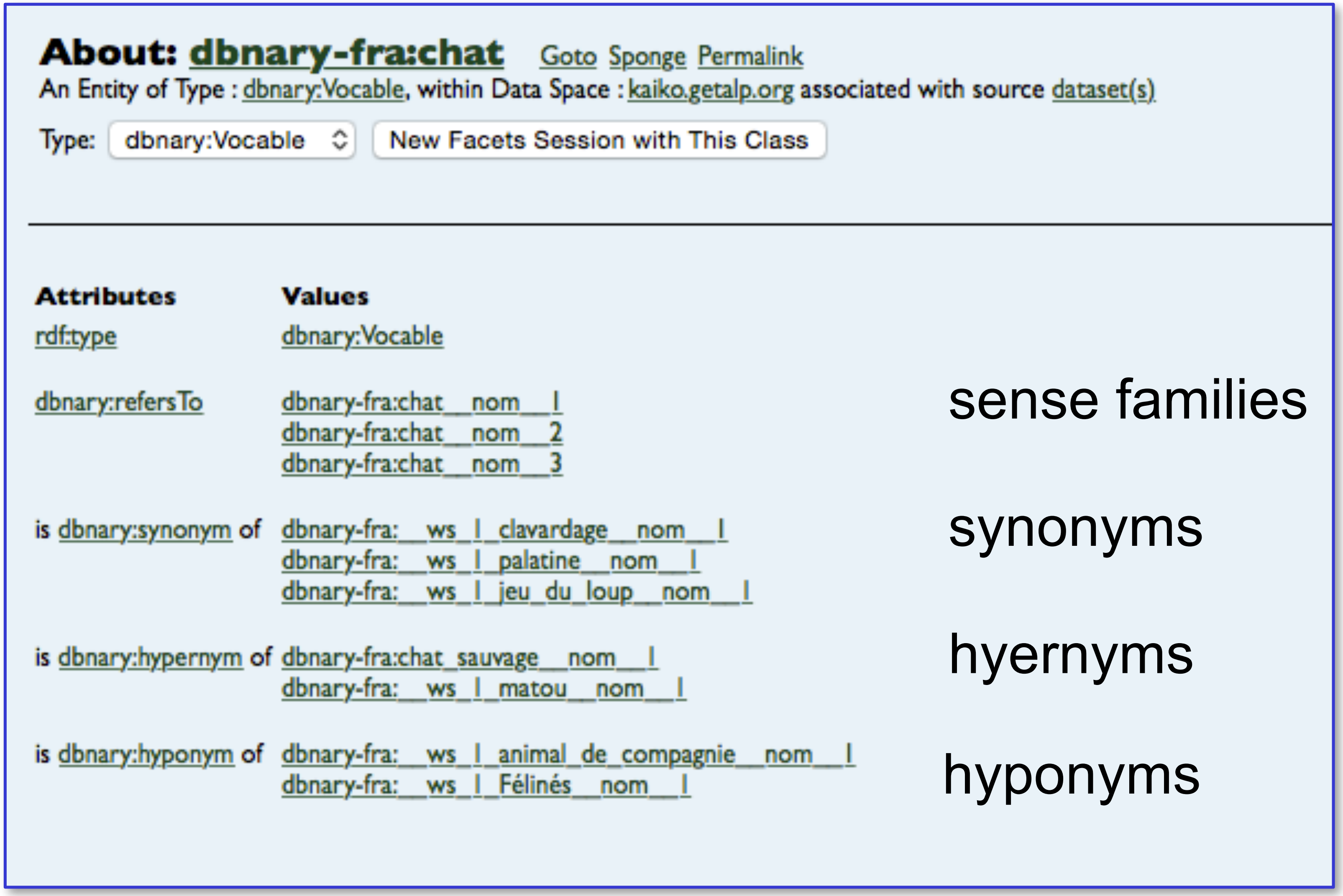


LEMON

A quick overview

Dbnary example: entry *chat* in French

- http://kaiko.getalp.org/dbnary/fra/chat



About: dbnary-fra:chat Goto Sponge Permalink

An Entity of Type : dbnary:Vocable, within Data Space : kaiko.getalp.org associated with source dataset(s)

Type: dbnary:Vocable ⇕ New Facets Session with This Class

Attributes	Values
rdf:type	dbnary:Vocable
dbnary:refersTo	dbnary-fra:chat__nom__1 dbnary-fra:chat__nom__2 dbnary-fra:chat__nom__3
is dbnary:synonym of	dbnary-fra:__ws_1_clavardage__nom__1 dbnary-fra:__ws_1_palatine__nom__1 dbnary-fra:__ws_1_jeu_du_loup__nom__1
is dbnary:hypernym of	dbnary-fra:chat_sauvage__nom__1 dbnary-fra:__ws_1_matou__nom__1
is dbnary:hyponym of	dbnary-fra:__ws_1_animal_de_compagnie__nom__1 dbnary-fra:__ws_1_Félinés__nom__1

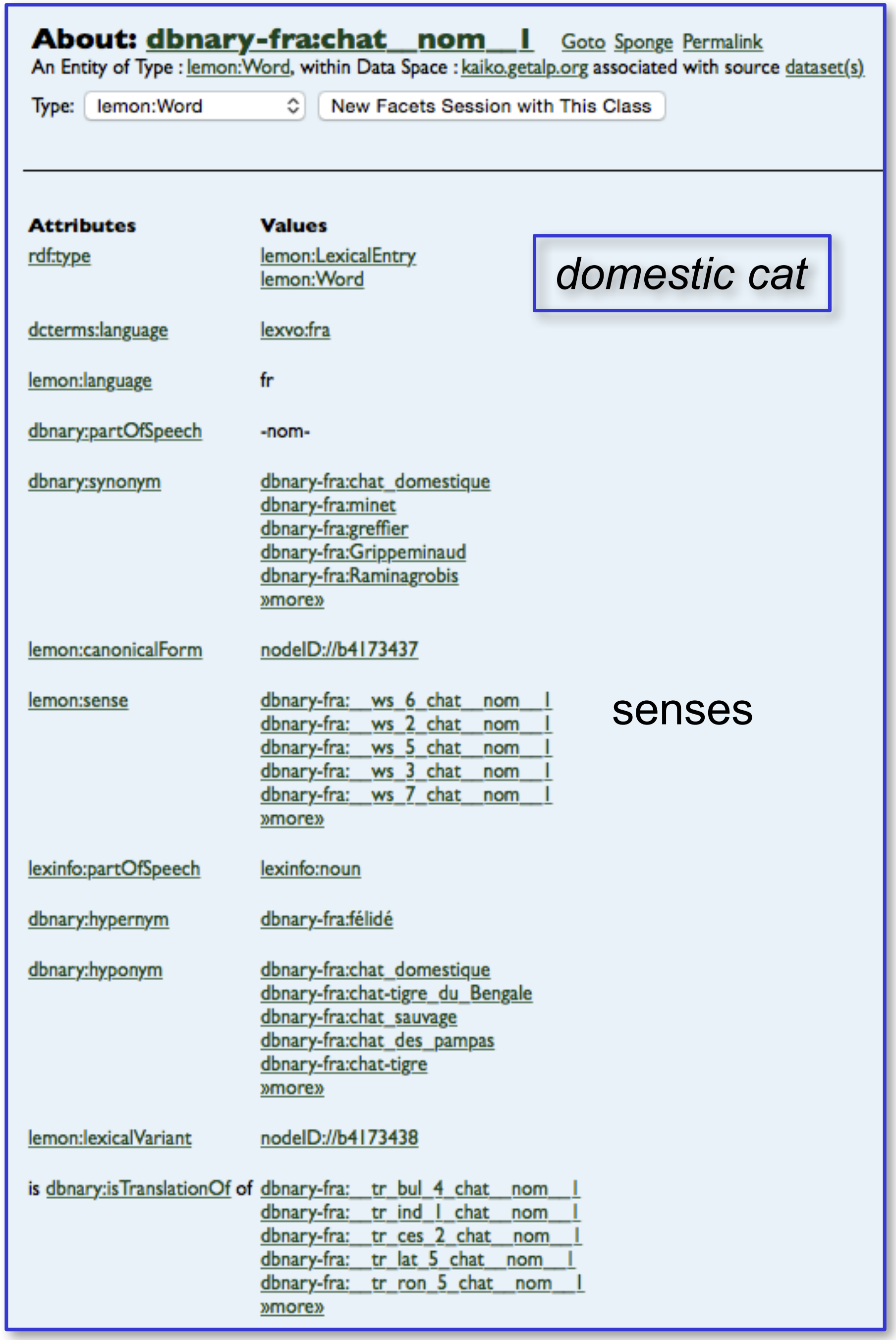
sense families

synonyms

hyernyms

hyponyms

fr chat → en cat; fr chat → en chat



About: dbnary-fra:chat__nom__1 Goto Sponge Permalink

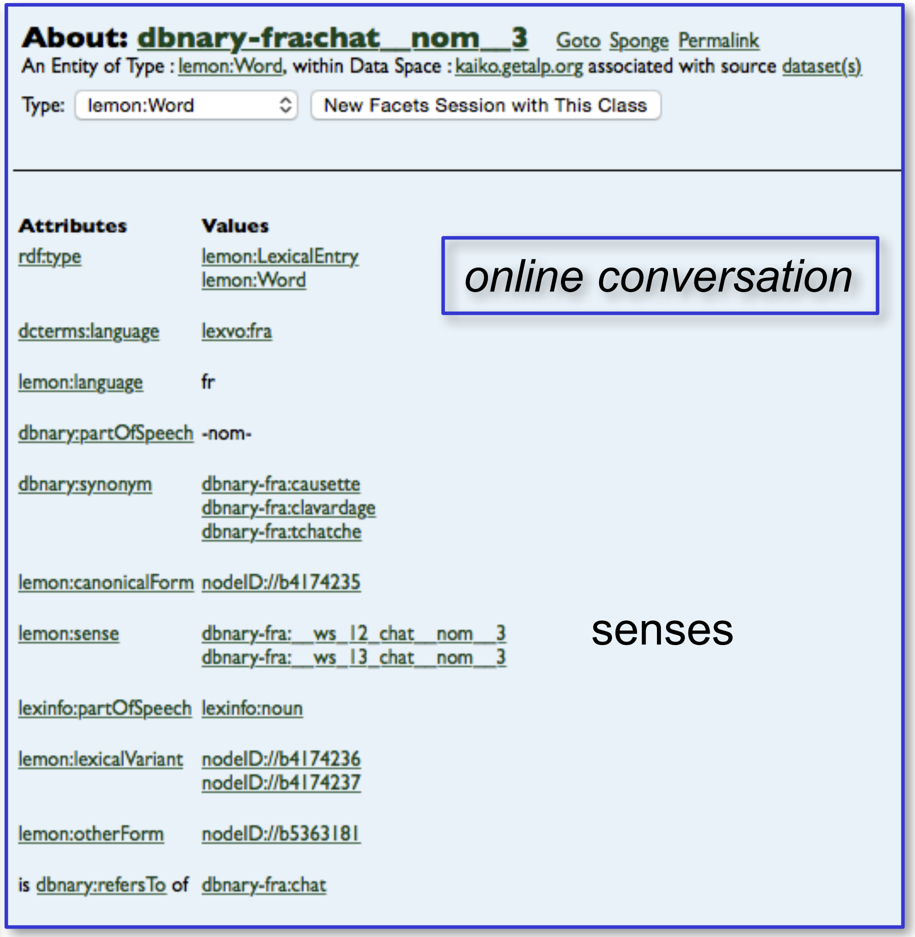
An Entity of Type : lemon:Word, within Data Space : kaiko.getalp.org associated with source dataset(s)

Type: lemon:Word ⇕ New Facets Session with This Class

Attributes	Values
rdf:type	lemon:LexicalEntry lemon:Word
dcterms:language	lexvo:fra
lemon:language	fr
dbnary:partOfSpeech	-nom-
dbnary:synonym	dbnary-fra:chat_domestique dbnary-fra:minet dbnary-fra:greffier dbnary-fra:Grippeminaud dbnary-fra:Raminagrobis »more»
lemon:canonicalForm	nodeID://b4173437
lemon:sense	dbnary-fra:__ws_6_chat__nom__1 dbnary-fra:__ws_2_chat__nom__1 dbnary-fra:__ws_5_chat__nom__1 dbnary-fra:__ws_3_chat__nom__1 dbnary-fra:__ws_7_chat__nom__1 »more»
lexinfo:partOfSpeech	lexinfo:noun
dbnary:hypernym	dbnary-fra:félidé
dbnary:hyponym	dbnary-fra:chat_domestique dbnary-fra:chat-tigre_du_Bengale dbnary-fra:chat_sauvage dbnary-fra:chat_des_pampas dbnary-fra:chat-tigre »more»
lemon:lexicalVariant	nodeID://b4173438
is dbnary:isTranslationOf of	dbnary-fra:__tr_bul_4_chat__nom__1 dbnary-fra:__tr_ind_1_chat__nom__1 dbnary-fra:__tr_ces_2_chat__nom__1 dbnary-fra:__tr_lat_5_chat__nom__1 dbnary-fra:__tr_ron_5_chat__nom__1 »more»

domestic cat

senses



About: dbnary-fra:chat__nom__3 Goto Sponge Permalink

An Entity of Type : lemon:Word, within Data Space : kaiko.getalp.org associated with source dataset(s)

Type: lemon:Word ⇕ New Facets Session with This Class

Attributes	Values
rdf:type	lemon:LexicalEntry lemon:Word
dcterms:language	lexvo:fra
lemon:language	fr
dbnary:partOfSpeech	-nom-
dbnary:synonym	dbnary-fra:causette dbnary-fra:clavardage dbnary-fra:tchatche
lemon:canonicalForm	nodeID://b4174235
lemon:sense	dbnary-fra:__ws_12_chat__nom__3 dbnary-fra:__ws_13_chat__nom__3
lexinfo:partOfSpeech	lexinfo:noun
lemon:lexicalVariant	nodeID://b4174236 nodeID://b4174237
lemon:otherForm	nodeID://b5363181
is dbnary:refersTo of	dbnary-fra:chat

online conversation

senses

Dbnary: a source of Synsets for METEOR?

- The big picture
  - 21 languages
  - 2.9M lexical entries (pos, canonical form, +{})
    - divided into 2.5M senses (def, example)
  - 4.9M translations (from 21 languages)
- We will consider the following languages

	English	French	Russian	German	Spanish
# of entries	620,369	322,018	185,910	104,505	86,388
# of senses	498,415	416,323	176,335	116,290	126,411
#of synonyms	35,437	36,019	31,345	33,282	21,024

Table 2. Number of entries, senses, and synonyms in Dbnary for the target languages considered in this study.

SYNSET EXTRACTION FROM DBNARY

Querying Dbnary

- SPARQL queries to extract every synonym (?s) in the Dbnary database for each word (?w) in a specific language

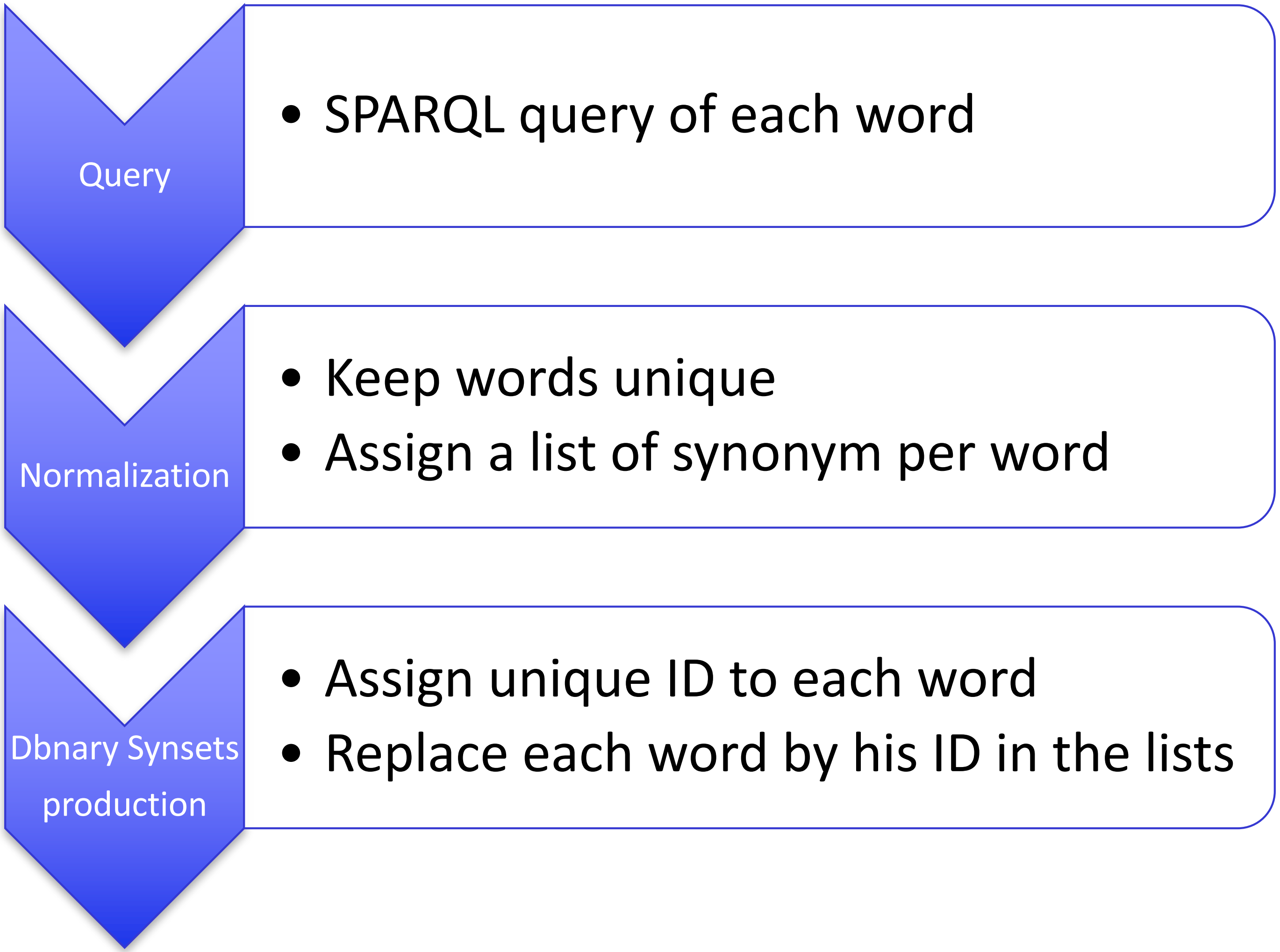
```
SELECT distinct ?w  ?s
WHERE {  ?s dbnary:synonym ?w.
         ?w dbnary:refersTo ?le.
         ?le lemon:language 'en'.}
```

- Example
  - ?w = "cut"

lower	reduce	juice	decrease
vigorish	decrease	ripped	cutting

Producing the Synsets

- Produce *a la* WordNet Synsets from Dbnary



Query

- SPARQL query of each word

Normalization

- Keep words unique
- Assign a list of synonym per word

Dbnary Synsets production

- Assign unique ID to each word
- Replace each word by his ID in the lists

2 dictionaries of synonyms

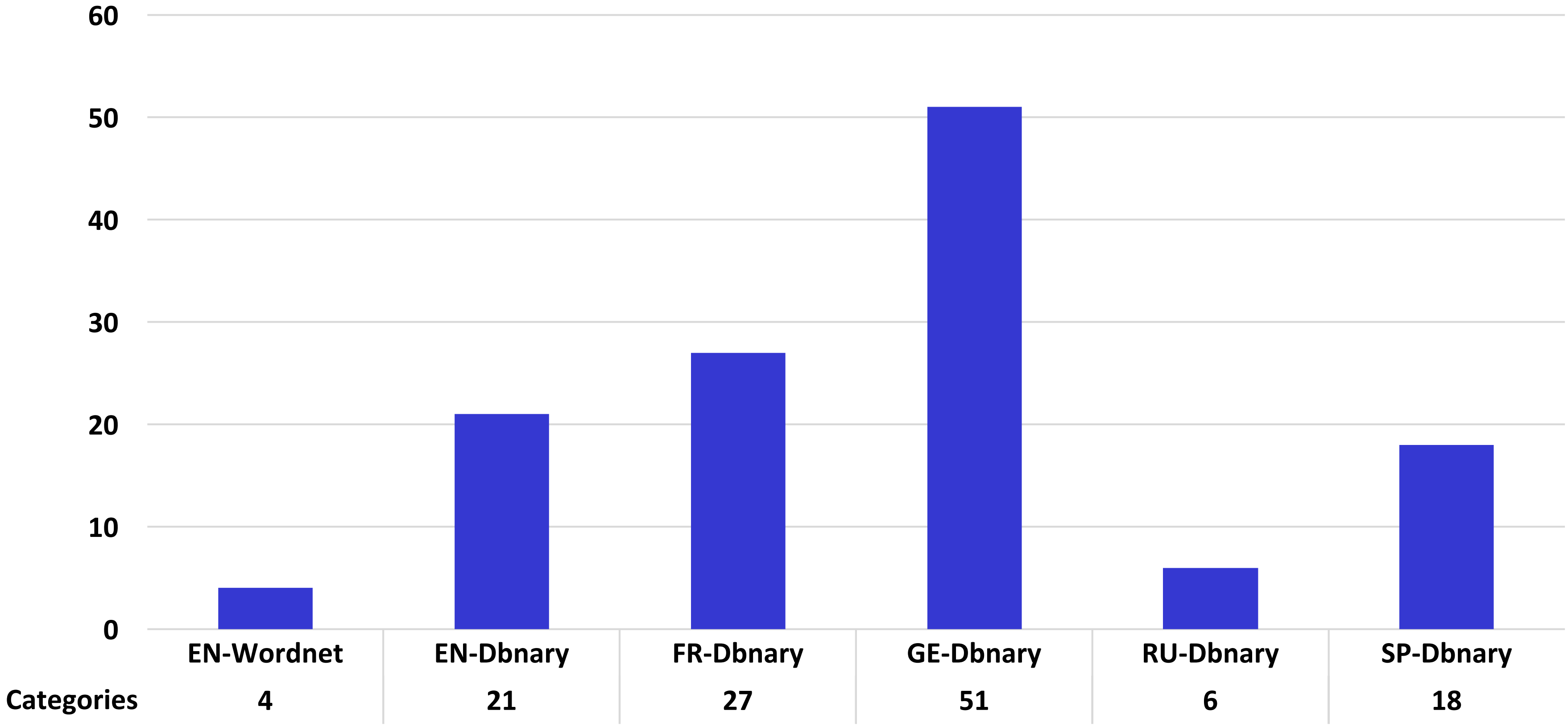
- **DB-4-catg**
  - with the 4 WordNet categories: **Verb, Noun, Adverb, Adjective**
- **DB-all-catg**
  - with all the existing categories in Dbnary

category		
Noun	Phrase	Proverb
Adjective	Suffix	Numeral
Verb	Pronoun	Determiner
Adverb	Prep_phr	Symbol
Proper_noun	Conj	Card_num
Interjection	Prefix	Infix
Preposition	Particle	Idiom

Table 3. All category extracted from Dbnary for English

# of Dbnary categories/language

- # of categories for the languages considered
  - English, French, German, Russian, and Spanish



	EN-Wordnet	EN-Dbnary	FR-Dbnary	GE-Dbnary	RU-Dbnary	SP-Dbnary
Categories	4	21	27	51	6	18

Scores comparison with reported results of WMT14 on French-**English**

- ✔ WordNet original Synsets (4 categories)
- ✔ “Dbnary 4 cats Synsets” (**DB-4-catg**)
- ✔ “Dbnary 21 cats Synsets” (**DB-all-catg**)

METEOR SCORES ON ENGLISH WORDNET VS DBNARY

Impact of the “Synsets”

METEOR	Baseline (WordNet)	DB-4-catg	DB-all-catg
online A	**36.97** %	**36.91** %	**37.13** %
rbmt-1	**33.74** %	**33.60** %	**33.89** %

Table 4 . METEOR-Baseline vs METEOR-Dbnary for 2 randomly picked up systems from WMT14 data (French-English MT)

- Comments
  - similar scores for the **Baseline** & **DB-4-catg**
    - the size of the WordNet dictionary is 2,5 times larger than the size of Dbnary (4-catg).
  - small increase (>0.2, >0.6%) using all 21 Dbnary categories with **DB-all-Catg**

The second hidden parameter

- METEOR uses the **Morphy-7WN** function to find the lemma of a given English word
  - what would we do for the other languages?
- Idea
  - Use Treetagger *(Schmid, 1994)* to lemmatize forms for any language
- Cons
  - Using Treetagger while computing METEOR score will slow down the execution time
- Solution
  - preprocess the data (hypo, ref) to get lists of pairs (word, lemma)

Impact of the lemmatizer

	METEOR-Morphy	METEOR-TTG
online A	36.97 %	37.00 %
rbmt-1	33.74 %	33.76 %

Table 5. Impact of lemmatization; METEOR-Morphy vs METEOR-TTG for 2 randomly picked up systems from WMT14 data (French-English MT)

- Comment
  - A slight increase between the scores of METEOR-Morphy and METEOR-TTG
  - Possible explanation
    - TreeTagger lemmatizes all categories
    - Morphy-7WN lemmatizes only three categories (Noun, Verb and Adjective)

Correlation comparison with previously reported results

- ✔ English–Spanish (WMT13)
- ✔ French–English, English–French, English–Russian, English–German (WMT14)

CORRELATION WITH HUMAN JUDGMENT METEOR WORDNET VS DBNARY

Goal

- Compare correlation of METEOR and METEOR-Dbnary with human judgments of MT hypotheses
  - WMT13 Metrics Shared Task *(Machacek and Bojar, 2013)*
    - English–Spanish
  - WMT14 Metrics Shared Task *(Machacek and Bojar, 2014)*
    - French–English, English–French, English–German and English–Russian
- Evaluation measures
  - System-level: Pearson correlation coefficient between system rankings based on human judgments *vs* automatic score
  - Segment-level: Kendall's τ rank correlation coefficient between system rankings based on human judgments *vs* automatic score

Setup

- “Dbnary Synsets” for all the target languages:
  FR, SP, RU, GE
  - weight of 0.8 for the synonyms for each language
    - same weight as the English synonym module in the METEOR default setting
- Two configurations of METEOR
  - **METEOR-Baseline**: METEOR Universal (v1.5) with the synonym module activated for English only with WordNet
  - **METEOR-Dbnary**: METEOR Universal with the synonym module activated for EN, FR, SP, RU, GE, using “Dbnary Synsets”

Results for Pearson Correlation Coeff.

	WMT14				WMT13
	FR-EN	EN-FR	EN-RU	EN-GE	EN-ES
Meteor-Baseline	**.975**	.941	.923	.263	.886
Meteor-Dbnary	.973	**.943**	**.928**	**.320**	**.895**

Table 6. System-level correlations (Pearson Correlation Coefficient) between Baseline (or METEOR-Dbnary) and the WMT13/WMT14 human rankings

Comments

- when WordNet Synsets are available (**FR–EN**)
  - slight degradation (size(Dbnary) << size(WordNet))
- when WordNet Synsets are not available (**EN–XX**)
  - use of “Dbnary Synsets” slightly improves system-level correlations of METEOR score with human judgment

Results for Kendall's τ rank corr. coeff.

	WMT14				WMT13
	FR-EN	EN-FR	EN-RU	EN-GE	EN-ES
Meteor-Baseline	.406	.280	.238	.427	.184
Meteor-Dbnary	.406	**.284**	**.240**	**.435**	**.187**

Table 7. Segment-level correlations (Kendall's τ) between METEOR-Baseline (or METEOR-Dbnary) and the the WMT13/WMT14 human rankings

- Comments
  - Same trend that before for segment-level correlations
    - Dbnary can be a useful resource for MT evaluation to bring synonyms as an added feature

Changes in the METEOR score

	WMT14			WMT13
	EN-FR	EN-RU	EN-GE	EN-ES
Meteor-Baseline	50.94	36.21	38.06	49.88
Meteor-Dbnary	**52.34**	**37.60**	**41.51**	**51.04**

Table 8 : Comparison of METEOR-Baseline without synonyms vs METEOR-Dbnary (for *rbmt-1* system)

- Comments
  - METEOR-Dbnary scores are better
- Explanation
  - Using Dbnary as a lexical resource for synonymy, the metric maps more words with the same meaning

Example 1

- **Reference**: […] alors les **dirigeants** d'entreprise sont sûrement **aussi** des cibles potentielles.
- **Hypothesis**: […] alors sûrement les **chefs** de file des affaires sont **également** les cibles potentielles.

Synonym match

Word	Lemma	Synonym list
dirigeants	dirigeant	[**chef**, maître, leader, directeur]
chefs	chef	[tête, maître, cuisinier, leader, maître_queux, patron]
aussi	aussi	[ainsi, **également**, itou]
également	également	[**aussi**, pareillement, de_même, par_ailleurs]

- **Segment score:**

METEOR-Baseline: 0.6762

METEOR-Dbnary : 0.7290

Example 2

- **Reference**: J'estime qu'il est concevable que ces données soient **utilisées** dans leur intérêt mutuel.
- **Hypothesis**: Je pense qu'il est concevable que ces données soient **employées** pour le bénéfice mutuel.

Synonym match

Word	Lemma	Synonym list
utilisées	utiliser	[user]
employés	employer	[occuper, utiliser]

- **Segment score :**

METEOR-Baseline : 0.6609

METEOR-Dbnary : 0.7133

Example 3

- **Reference**: Il me parlait, m’encourageait constamment, il **habitait** mon corps.
- **Hypothesis**: Il me parlerait, m'encouragent constamment, il a **vécu** dans mon corps.

Synonym match

Word	Lemma	Synonym list
habitait	habiter	[occuper]
vécu	vivre	[habiter, nourriture]

- **Segment score :**

METEOR-Baseline : 0.6743

METEOR-Dbnary : 0.7688

OBJECTIVE EVALUATION

FINAL REMARKS

Pros of objective evaluation

- Costless
  - No! References have to be produced at some point!
- Objective
  - OK, always the same results with the same hypo & ref(s)
- Reusable
  - Always on the same test set (not a real life situation)
  - Correlation between "translation improvement" & "score improvement"

- System optimization
  - *is it good or bad?*
- System comparison
  - *as far as they use the same development protocol! (cf. IWSLT 04)*

Cons of objective evaluation

- System over tuning
  - When system parameters are adjusted towards the main evaluation metric
    - if it is BLEU then tune with BLEU, if it is NIST then tune with NIST
  - Several metrics used for ranking
- Blind system development
  - When metrics are unable to capture system improvements
- Unfair system comparison
  - When metrics are unable to reflect difference in quality between MT systems
  - When systems are based on different paradigms (SMT vs. RBMT) *(cf. IWSLT 2004)*
- No utility, usability evaluation yet

CONCLUSION

To be remembered

- On BLEU [Callison-Burch et al., 2006]
  - Under some circumstances an improvement in BLEU is *not sufficient* to reflect a genuine improvement in translation quality
  - Under other circumstances that it is *not necessary* to improve BLEU in order to achieve a noticeable improvement in translation quality
- To be transposed to all other objective metrics!

External vs internal measures

- External measures
  - linguistic criteria: grammaticality, fidelity…
  - usage criteria: productivity, cost, delay…
  - conflict between linguistic & usage criteria
    - ex: Systran, Euratom, ISPRA: 2/20 (linguistic quality) — 18/20 (usability)
- Internal measures
  - system design: linguistic & computational architecture
  - perspectives of improvements: quality, coverage
  - ease of extension to
    - new languages
    - new document types
    - new tasks (assimilation ⟶ dissemination)

Classification of external measures

- Measures related to the task
  - High quality written communication

    two tasks: acquisition (from one language source), diffusion (to one target language)
    - Produce a professional quality translation
    - ✧ reduction of costs (human labor) and delays
  - Spoken communication
    - Help two people to conduct a bilingual dialogue to accomplish a task
    - ✧ accomplishment of the task
  - Comprehension, understanding of written material
    - Translate Web pages, newspapers, e-commerce services so that end users can understand information in foreign languages and act accordingly
    - ✧ number of purchases per visited page in e-commerce, time spent reading newspapers page (objectives measures)
    - ✧ user feedback, answers to customer questionnaires (subjective measures)

Classification of external measures

Measures related to the task *(cont.)*

- Comprehension, understanding of spoken material

  the typical task is to follow a monologue (speech, Parliament, etc.). or a dialogue in a foreign language (television, intelligence)

  - Produce as much information as possible
  - determine the level of understanding
    - objective measure: time to complete the task, MCQ about the content
    - subjective measures: sense of understanding, judgment of fluidity

Classification of external measures

- Measures non related to the task
  - with references
    - adequacy *a la* NIST
    - fidelity *a la* JEIDA or FEMTI
    - informativeness *a la* ALPAC
  - without references
    - fluidity *a la* NIST
    - adequacy through MCGQ *a la* TOEFL or TOEIC

Proposal

Use only cheap task-related measures for external evaluation!

MT for written input

Diffusion

- objective usability measures
  - time spend for post-edition, correction of raw MT output
  - **Relative Efficiency**:

$$\text{Relative Efficiency}_{MT} = \frac{\text{Time}_{Human}}{\text{Time}_{MT+Human}}$$

- an MT system may be considered efficient if it's relative efficiency is > 2 (upper bound of the gain with a translation memory)

- subjective measure such as fluency or adequacy are useless and counterproductive
  - corrections made easy by the environment *(cf. "is admission fee how much?"*

Proposal

MT for written input

Acquisition, understanding

- Web pages
  - compare reading time translated Web page vs reading time original Web page
    - if shorter: very bad translation
    - if longer: bad translation but usable for some understanding
    - if equal: quality OK of the use
  - Multiple Choice Questions

Proposal

- **MT for spoken input**
  - **Diffusion**
    - MCQ for understanding
  - **Acquisition, Understanding**
    - MCQ but hard for dialogue

Final words…

- External methods for evaluating MT systems define various measures based on MT results and their usage.
- While operational systems are mostly evaluated since long by task-based methods, evaluation campaigns of the last years use (parsimoniously) quite expensive subjective methods based on unreliable human judgments, and (for the most part) methods based on reference translations, that are impossible to use during the real usage of a system, less correlated with human judgments when quality increases, and totally unrealistic in that they force to measure progress on fixed corpora, endlessly retranslated, and not on new texts to be translated for real needs.
- There are also numerous biases introduced by the desire to diminish costs, in particular the usage of parallel corpora in the direction opposed to that of their production, and of monolingual rather than bilingual judges.
- We propose to abandon the reference-based methods in external evaluations, and to replace them by strictly task-based methods, while reserving them for internal evaluations.

BIBLIOGRAPHY

Bibliography (1/5)

- ALPAC (1966). *Language and Machine: Computers in Translation and Linguistics.* n. 1416. Automatic Language Processing Advisory Committee, Division of Behavioral Sciences, National Academy of Science - National Research Council. Washington, D. C. November 1966. 138 p.
- Babych, B., & Hartley, A. (2004). *Extending BLEU MT Evaluation Method with Frequency Weightings*. Proceedings of ACL 2004. Barcelona, Spain. July 21-26, 2004. pp. 622-629.
- Banerjee, S., & Lavie, A. (2005). *METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgement*. Proceedings of ACL-05, Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. Ann Arbor, USA. June 29, 2005. pp. 25-32.
- Blanchon, H. (2004). *HLT Modules Scalability within the NESPOLE! Project*. Proceedings of ICSLP. Jeju Island, Korea. October 4-8, 2004. 4 p.
- Blanchon, H., Boitet, C., & Besacier, L. (2004). *Spoken Dialogue Translation System Evaluation: Results, New Trends, Problems and Proposals*. Proceedings of IWSLT 2004. Kyoto, Japan. September 30 - October 1, 2004. pp. 95-102.

Bibliography (2/5)

- Blanchon, H., Boitet, C., Brunet-Manquat, F., Tomokio, M., Hamon, A., Hung, V. T. et al. (2004). *Towards Fairer Evaluation of Commercial MT Systems on Basic Travel Expressions Corpora*. Proceedings of IWSLT 2004. Kyoto, Japan. September 30 - October 1, 2004. pp. 21-26.
- Callison-Burch, C., Osborne, M., & Koehn, P. (2006). *Re-evaluating the Role of BLEU in Machine Translation Research*. Proceedings of ACL-2006. Trento, Italy. April 3-7, 2006. pp. 249-256.
- Doddington, G. (2002). *Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics*. Proceedings of HLT 2002. San Diego, California. March 24-27, 2002. pp. 138-145.
- EAGLES-EWG (1996). *EAGLES – Evaluation of Natural Language Processing Systems.* Final Report EAG-EWG-PR.2, Project LRE-61-100. Center for Sprogteknologi. Copenhagen, Denmark. October, 1996. 287 p.
- EAGLES-EWG (1999). *EAGLES – Evaluation Working Group.* Final Report EAG-II-EWG-PR.2, Project LRE-61-100. Center for Sprogteknologi. Copenhagen, Denmark. April, 1999. 173 p.
- Hovy, E., King, M., & Popescu-Belis, A. (2002). Principles of Context-Based Machine Translation Evaluation. *Machine Translation*, 17(1): pp. 43-75.

Bibliography (3/5)

- JEIDA (1989). *A Japanese view of machine translation in light of the considerations and recommendations reported by ALPAC, U.S.A.* Japan Electronic Industry Development Association. Tokyo, Japan. July, 1989. 197 p.
- JEIDA (1992). *JEIDA Methodology and Criteria on Machine Translation Evaluation.* Japan Electronic Industry Development Association. Tokyo, Japan. November, 1992. 129 p.
- King, M., Popescu-Belis, A., & Hovy, E. (2003). *FEMTI: creating and using a framework for MT evaluation*. Proceedings of MT Summit IX. New Orleans, USA. September 23-27, 2003. 8 p.
- Leusch, G., Ueffing, N., & Ney, H. (2006). *CDER: Efficient MT evaluation using block movements*. Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento, Italy. April 3-7, 2006. pp. 241-248.
- Lin, C.-Y., & Och, F. J. (2004). *Automatic Evaluation of Machine Translation Quality Using Longest Common Subsequence and Skip-Bigram Statistics*. Proceedings of ACL 2004. Barcelona, Spain. July 21-26, 2004. pp. 605-612.

Bibliography (4/5)

- Lita, L. V., Rogati, M., & Lavie, A. (2005). *BLANC: learning evaluation metrics for MT*. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, B.C., Canada. October 6-8, 2005. pp. 740-747.
- Liu, D., & Gildea, D. (2006). *Stochastic Iterative Alignment for Machine Translation Evaluation*. Proceedings of COLING-ACL. Sydney, Australia. 17-21 July, 2006. pp. 539-546.
- Melamed, I. D., Green, R., & Turian, J. P. (2003). *Precision and Recall of Machine Translation*. Proceedings of HLT-NAACL 2003 - short papers. Edmonton, Canada. May 27 - June 1, 2003. pp. 61–63.
- Nießen, S., Och, F. J., Leusch, G., & Ney, H. (2000). *An Evaluation Tool for Machine Translation: Fast Evaluation for MT Research*. Proceedings of LREC 2000. Athens, Greece. 31 May - 2 June, 2000. pp. 39-45.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). *BLEU: a Method for Automatic Evaluation of Machine Translation*. Proceedings of ACL-02. Philadelphia, USA. July 7-12, 2002. pp. 311-318.

Bibliography (5/5)

- Przybocki, M., Sanders, G., & Le, A. (2006). *Edit Distance: A Metric for Machine Translation Evaluation*. Proceedings of LREC 2006. Genoa, Italy. May 24-26, 2006. pp. 2038-2043.
- Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). *A Study of Translation Edit Rate with Targeted Human Annotation*. Proceedings of AMTA 2006. Cambridge, MA, USA. August 8-12, 2006. pp. 223-231.
- Tillmann, C., Vogel, S., Ney, H., Zubiaga, A., & Sawaf, H. (1997). *Accelerated DP-based search for statistical translation*. Proceedings of Fifth European Conference on Speech Communication and Technology (EUROSPEECH'97). Rhodos, Greece. September 22-25, 1997. pp. 2667-2670.
- Turian, J. P., Shen, L., & Melamed, I. D. (2003). *Evaluation of Machine Translation and its Evaluation*. Proceedings of MT Summit IX. New Orleans, USA. September 23-27, 2003. pp. 386-393.
- White, J. S., O'Connell, T., & O'Mara, F. E. (1994). *The ARPA MT Evaluation Methodologies: Evolution, Lessons and Further Approaches*. Proceedings of Technology Partnerships for Crossing the Language Barrier (the First Conference of the Association for Machine Translation in the Americas). Columbia, Maryland, USA. October 5-8, 1994. pp. 193-205.