linguatec
Sprachtechnologien

Automatic Means of MT Evaluation

Gr. Thurmair

ELRA HLT Evaluation Workshop
MALTA, Dec. 2005

Latest German MT evaluation campaign

Übersetzungssoftware – Checkliste

	@promt	Hexaglot Volltext-übersetzer 5.0	Langenscheidt T1 Englisch 5.0	Personal Translator 2006	PowerTranslator 9	translate 8
Hersteller	Promt	Hexaglot	Langenscheidt	linguatec	LEC	Lingenio
Vertrieb					Avanquest	digital publishing
Web-Adresse	www.promt.de	www.hexaglot.de	www.langenscheidt.de	www.linguatec.de	www.avanquest.com/de	www.lingenio.de
Registrierung erforderlich	–	–	–	✓	✓	–
Lieferumfang						
Dokumentation gedruckt/auf CD	Handbuch (360 S.)/✓	Booklet (14. S.)/ PDF (205 S.)	–/PDF (147 S.)	Handbuch (117 S.)/✓	Booklet(83 S.)/✓	Handbuch (144 S.)/–
Satzarchiv	✓	–	✓	✓	–	✓
Add-ins für Word/Excel/PowerPoint/Outlook	✓/✓/✓/✓	✓/✓/✓/✓	✓/✓/✓/✓	✓/✓/✓/✓	✓/✓/✓/–	✓/✓/✓/✓
weitere Add-ins	MS FrontPage, Adobe Reader	Internet Explorer	Internet Explorer, MS Access	Adobe Reader	Internet Explorer	–
Sprachausgabe/OCR-Software	✓/✓	–/–	–/✓	✓/–	–/–	–/–
Rechtschreibkontrolle/-vorschläge	✓/✓	–/–	–/–	✓/✓	–/–	✓/–
Wörterbücher						
Nachschlagen System-WB/flektierte Formen/per Mausklick	✓/✓/✓	–/–/–	✓/✓/✓	✓/✓/✓[1]	✓/✓/–	✓/✓/✓[1]
Idiomatikwörterbuch	✓	–	–[2]	✓	–	✓
Benutzerwörterbuch/mehrere möglich	✓/✓	✓/–	✓/–	✓/✓	✓/✓	✓/✓
Alternativeinträge möglich	✓	–	–	✓	–	✓
Import von Wortlisten	–	–	✓[2]	✓	–	✓
Verarbeitbare Formate						
Lesen: TXT/RTF/DOC/HTML/PDF	✓/✓/✓/✓/✓	✓/✓/✓/✓/–	✓/✓/–/✓/–	✓/✓/✓/✓/✓	✓/✓/✓/✓/–	✓/✓/✓/✓/–
Schreiben: TXT/RTF/DOC/HTML/PDF	✓/✓/✓/✓/✓	–/–/✓/–/–	✓/✓/–/✓/–	✓/✓/✓/✓/✓	✓/✓/✓/✓/–	✓/✓/✓/✓/–
Bilddaten (BMP, TIF, JPG)	✓	–	✓[3]	–	–	–
Import/Export von TMX-Daten	✓	–	✓	✓	–	✓
Bedienung						
kontextsensitive Hilfe	✓	–	–	✓	–	✓
Editor/zwei Texte horizontal/vertikal	✓/✓/✓	–/–/–	✓/–/✓	✓/✓/✓	–/✓/–	✓/✓/–
Wörter ausnehmen per Markierung	✓	–	✓	✓	–	✓
Übersetzen in Word: ab Einfügemarke/Markierung/Absatz	✓/✓/✓	–/–/–	–/✓/–	✓/✓/✓	–/✓/–	✓/✓/✓
Übersetzen in Excel: Markierung/Blatt/Mappe	✓/✓/✓	–/–/✓	✓/–/✓	✓/✓/–	✓/✓/–	✓/✓/–
Übersetzen in PowerPoint: Markierung/Folie/Präsentation	✓/✓/✓	–/✓/✓	–/✓/✓	✓/✓/✓	✓/–/✓	–/✓/✓
Zeitbegrenzung einstellbar	–	–	–	–	–	✓
Speichern Quelle/Ziel/beides/Tabellenform	✓/✓/✓/✓	✓/✓/–/–/–	✓/✓/–/–/–	✓/✓/–/–/–	✓/✓/–/–/–	✓/✓/–/–/–
gesamtes Projekt speichern/Backup-Funktion/Stapelverarbeitung	✓/✓/✓	–/–/–	–/–/–	✓/–/–	–/–/✓[4]	–/–/–
Übersetzung						
automatische Erkennung der Sprachrichtung	✓	–	–	✓	–	✓
Sachgebiet: manuelle Wahl/autom. Erkennung	✓/✓	–/–	✓/–	✓/✓	–/–	✓/–
Eigennamen: Kennzeichnung möglich/autom. Erkennung	✓/–	✓/–	✓/–	✓/✓	–/–	✓/–
Synchronisier-Funktion	✓	–	✓[2]	✓	–	✓
Wahlmöglichkeit bei Pronomina/Anrede	✓/✓	–/–	–/✓	✓/✓	–/✓	✓/✓
dt. Quelltext: Wahlmöglichkeit bei Komposita/alte oder neue RS	✓/✓	–/–	–/✓	✓/✓	–/–	✓/✓
engl. Quelltext: Wahlmöglichkeit bei Imperativ/UK- oder US-Englisch	–/✓	–/✓	✓/✓	✓/✓	–/–	✓/✓
unterschiedliche Markierungen für Alternativen/unbek. Wörter/ Eigennamen/Komposita im Zieltext	✓/✓/✓/✓	–/✓/–/–	✓/✓/✓/✓	–/✓/✓/–	–/–/–/–	–/✓/–/–
unbekannte Wörter: suchen/Liste drucken	✓/✓	–/–	✓/✓	✓/–	–/–	✓/–
Zeilenumbruch als Satzende wahlweise	✓	–	–	✓	–	✓
Bewertung						
Dokumentation	○	⊖	○	⊕	⊖	⊕
Bedienung	○	○	⊕	⊕	○	⊕⊕
Integration in andere Anwendungen	⊕⊕	⊕⊕	○	⊕	⊖	⊕
Übersetzung Alltagstexte	⊕⊕	⊕	⊕	⊕⊕	⊕	⊕⊕
Übersetzung technische/wissenschaftliche Texte	⊕⊕	○	○	⊕	⊖	⊕
Wortschatz	⊕⊕	⊖	⊕	⊕⊕	○	⊕⊕
Preise						
getestete Version	Professional 300 €	Office 60 €	Professional 300 €	Professional 250 €	Standard 50 €	translate pro 250 €
weitere	Office 100 € Expert 600 €		Standard 50 € Standard plus 100 €	Home 50 € Office 100 €	Professional 200 € (Engl., Frz., Italienisch)	translate 50 € translate plus 100 €

[1] nur aus Quelltext [2] nur Satzarchiv [3] OCR-Software muss separat installiert werden [4] verzeichnisweise

⊕⊕ sehr gut ⊕ gut ○ zufriedenstellend ⊖ schlecht ⊖⊖ sehr schlecht ✓ vorhanden – nicht vorhanden k. A. keine Angabe

Some Evaluation Criteria (c't 8/2005)

- Tools
  - Translation Memory
  - Addins for MS-Office / others
  - Speech Output / OCR input
  - Spell Checker
- Dictionaries
  - Lookup by mouse click
  - Lookup of inflected forms
  - Idiom dictionary
  - (several) user dictionaries
  - Import of word lists
- Formats
  - Doc, html, rtf, pdf, txt, tmx
- Handling
  - Contextual help
  - Editor, no-translate mark
  - Store source, target, project
- Translation
  - Language identification
  - Subject area recognition
  - Proper name recognition
  - Pronoun translation
  - Locale (us, uk)
  - Synchronisation
  - Markups for alternatives, unknowns, proper names
- Evaluation
- Price

=> A case for FEMTI!

Task of MT Evaluation

- FEMTI: Framework for MT Evaluation
  - Taxonomy of User Needs
    - User, task, nature of input
  - Taxonomy of System Characteristics
    - System features

  Merit: MT Evaluation can be located in a global setting
    - Several examples (Miller, Popescu, and others)

MT as specific kind of software

- FEMTI top down aproach: MT is *Software*
  - Evaluation for standard characteristics of software
    - ISO 9126: functionality – reliability – usability – efficiency – maintainability – portability
    - MT System design must support these criteria: =>
      tests for robustness, speed. ergonomic aspects, GUI, document format support, parameters, settings, …

      This is expected from commercial software
    - But it is NOT the hot issue for MT Evaluation

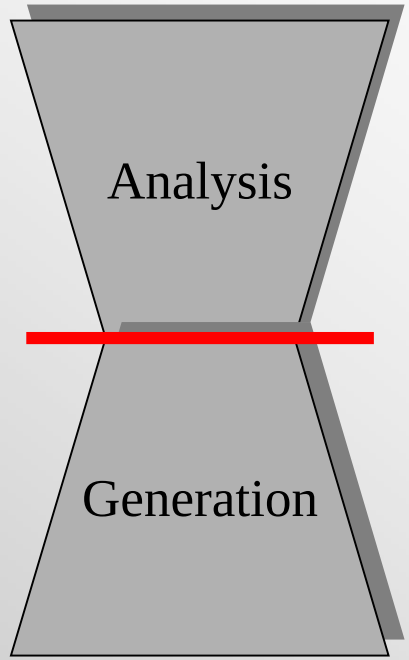
- Special functionality of MT software: Translation
  - Task of Translation:
    - **carry the *meaning*** to another language
    - (also: carry the *style* to another language, in literary translation)
  - MT Evaluation means: how well is this done?

Meaning preservation

- A formal evaluation needs a formal way of meaning representation! How to do this?

  "There is no good linguistic model of translation adequacy which can be easily formalised" (Babych)

  - Meaning representation as **interface**
    - Eurotra Interface Structure, MIR, QLF, UNL,,
    - Must be 'language-independent' / interlingual
    - Must cover all relevant aspects of meaning
      - Tense, aspect; definiteness, …
    - Faces the problem of lexical equivalences, …
  - Describe all **legal variants** on source and target
    - All legal mappings from input to representation
    - All legal mappings from there into target strings
  - This would enable a formal evaluation
    - Has a translation (variant) the same meaning?
  - But is far apart …



Evaluation techniques

- Until then: Meaning preservation evaluated by **human intuition**
  - Sets of evaluators
  - Scales of ratings, …
- Meaning preservation measured by evaluation **criteria**
  - Fidelity, informativeness, fluency, …
  - (these criteria need to be *related* properly to each other)
- Can this be optimised?
  - Find **automatic means** which have the same *results* than human intuition (but may use different heuristics / evaluation criteria)
  - But: make sure that the criteria *make sense* in themselves
- Automatic evaluation should be:
  - fast
  - objective
  - replicable
  - Automatic
  - cheap

Classical evaluation criteria

- Fidelity: evaluates the job of translation proper
  - Fidelity = "degree to which the information contained in the original text has been reproduced without distortion in the translation" (FEMTI)
  - (Criterium of Informativeness is a *consequence* of this)

- Fluency: holds for *any* text generation
  - Text generation must master the language it uses. Holds for
    - Language Learning (Reeder)
    - Automatic Abstracting / Text generation (A. Belz on NLG)
    - Text production and editing
    - Translation
  - Fluency is basically a **monolingual** job!
  - In MT it is only a supportive measure, used because MT is poor
    - Human translations do not show this problem
    - No apriori connection to MT (Candide more fluent than human, Babych)
    - It is easy to produce fluent non-content text (White)

(On fluency)

- As fluency is monolingual:
  Evaluate it on a monolingual basis:
  - By using language checking procedures
    - Spell checkers, grammar checkers
    - Lexical coherence, lexical verification
    - Language models,
    - **x-score** (compares document syntax to corpus syntactic structures)

- Fluency is related to fidelity
  - Translations can be fluent but inadequate
    - Wrong but more fluent lexical selection
    - BLEU brevity penalty reflects this
  - Some fluency scores in fact measure fidelity
    - E.g. if fluency is used to evaluate erratic lexical transfers
      - De Strom -> En current / river
        Stromunternehmen -> river enterprise: ??

Automatic evaluation of fidelity

Automatic evaluation means must be **bilingual**!

- because the *translation task* (meaning preservation) is bilingual!

Three main approaches

1. Mainstream technique
   - Test suites, human evaluation
2. Feature-based evaluation
   - Identify features which relate to fidelity in source and target
   - Correlate them
3. Document similarity evaluation
   - Create reference translation
   - Compare test texts with such references
   - Compute similarity

1. Mainstream technology

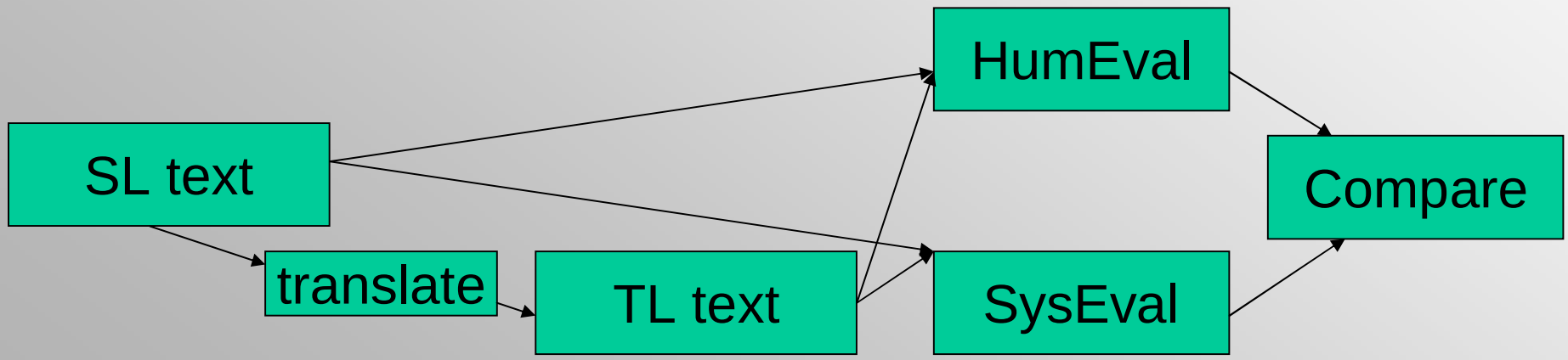
- Test Suites, consisting of
  - Test suites for **special phenomena** which the systems should cover
    - E.g.: pronouns; names; indirect speech;
    - Different translations for different system settings (subject area)
    - Databases with sentences containing such phenomena
      - TSNLP, 863-project (Liu)
  - **Real life texts** to check overall performance
    - Several K sentences

  Classification of translation pairs (on a 3-4 point scale)

- Evaluation by comparing new output with previous attempts
  - Run *diff* tools to identify differences
  - Evaluate differences in translation by hand (3-4 point scale)
    - Look at every different sentence pair …

2. Feature-based evaluation

- Identify features which can be compared in source and target, hope they support meaning preservation
  - Should be as language-independent as possible
  - Should be easily identifiable (cheap to produce)
- Extrapolate from them to the quality of the whole system
- (Compare automatic evaluation result with human judgement)



Feature-based evaluation

Such features can be taken from diferent text levels:

- String level
  - Used in text alignment (digits, relative segment length) (not used in MT)
  - Relevance of upper/lower case information (Turian)
- Morphological and dictionary level
  - Not really proposed (e.g.: SL plural / past -> TL plural / past)
  - May work for very close languages (Es – Ca)
  - But: lexicon coverage (general and special terminology) is an important feature for system comparison
    - **Unknown words** always block meaning preservation

Feature-based evaluation

- Syntactic level
  - Noun phrases (A de B -> B A or A of B) (White 2001)
  - Number of NPs / VPs: could be similar in source and target
    - (This is language dependent (compounds in De vs. NPs in En)
  - 'interlingual' X-score (syntactic dependencies)
    - Create and compare astract head-modifier structures in source and target language
- Plausibility of syntactic level features:
  - Constituent structure is too language specific
  - Dependency relations are much more stable
    - They describe *relations between concepts* on an abstract level
    - Positive experience from IR domain (TREC)
  - But: Evaluation result depends on the quality of analysis tools

Feature-based evaluation

- Semantic level
  - Named entities: must all show up in target (…)
  - S-score: position of document in vector space
  - Similar conceptual distances in SL and TL wrt WordNet
  - Compare predicate-argument situation in SL and TL (Belvin)
  - Similarities of conceptual clusters of lexical entries in SL and TL

- Plausibility of semantic level evaluation
  - Intuitive, content-wise
  - WordNet distances not very reliable (St.Onge)
  - S-score needs further investigation
    - (Divergent results reported)
  - More on evaluating good *transfer term selection* could be done

Example: Named Entity

? ?

MT 1: China halts German ambassador MA3 CAN4 RONG2 and legation staff member etc. also goes to airport to greet chairman Hu Jintao.

MT 2: Chinese Ambassador to Germany Ma Can Rong and embassy personnel etc. welcomes Chairman Hu Jin Tao also to the airport.

MT 3: Chinese Ambassador's to Germany horse's brilliant honour and embassy personnel,etc. go to the airport to meet President Hu Jintao.

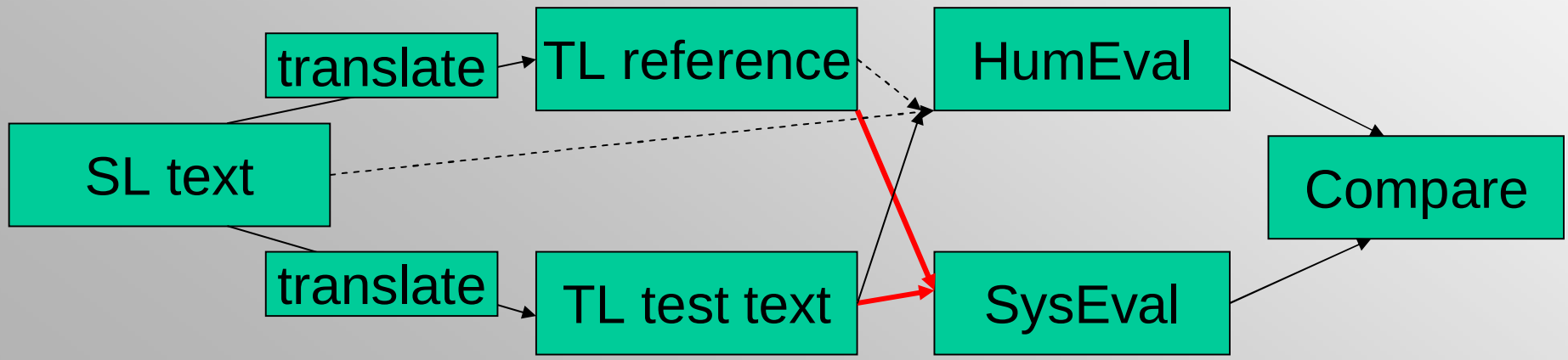
(Missing the name gives a bad translation.

Where is the named entity in the source? …)

3. Document similarity evaluation

- Identify / create documents with the same meaning
  - i.e. provide *reference translations*
- Compare the test translation results with this reference
  - Most similar document scores best

“The closer a machine translation is to a professional human translation, the better it is” (Papineni 2002) -> what is ‘closer’?



Evaluation results

- Total divergence on such metrics in literature
- Total divergence on inter-rater reliability

An automatic metric ideally predicts human judgement robustly across the spectrum of translation quality and across language families. … The BLEU score correlates highly with human judgements. (Papineni 2002)

The average inter-rater agreement correlation) is strong (Coughlin 2003) / Joint agreement at approximately 0.8, a respectable value (Miller 2005)

Even though human evaluation of MT is itself inconsistent and not very reliable, automatic MT evaluation measures are even less reliable and are still very far from being able to replace human judgement. (Turian et al. 2003)

It is obvious that neither the BLEU nor the NIST metric has the same effect as the human judgement (Zhang 2004)

=> Why is this so? What are determining factors of this divergence?
Possible Answers:

a. the **setup of the scenarios** creates strange effects,
b. the **content of the metrics** makes the results unstable

Possibility 1: Scenario Setup

- Reasons why automatic metrics do / do not match human judgement
  - In some settings human and automatic judgement are in alignment, in others they are not
  - Can be due to any component of the setup
    - Influence of source text?
    - Influence of reference translations?
    - Influencs of test translations?
    - Influence of human ratings?

Source Text?

- Influence of Source text
  - results depends on text type / corpus
    - Different for Microsoft TechManuals and Hansard general text (Coughlan 2003)
  - results depends on languages involved
    - Same setup, different results for Ar and En > Fr (Sourcin 2005) or De or Zh > En (Leusch 2003)
  - result depends on size of corpus
    - NIST: must be constant (Coughlin 2003)
    - BLEU: must be > 500 sentences
  - result depends on the size of doucments
    - Inter-judge correlation is low on shorter documents (Tourian 2003)
    - Human evaluation is less stable on longer documents

Depending on these paramaters, scores are reported to match / not to match human judgments

Reference Translations?

- **All scores crucially depend on the reference!**
- But: There is not *the* reference …

  “There is no one right translation of even a banal text … There just is no gold standard“ (King 2003)
- Reference translation is not gold standard production
  - A good translation is a free translation
  - But: more literal translations score better

    “The only professional translator got worse scores than the translations of all seven non-professionals … This is because the non-professional translations tended to be fairly literal and stayed as close to the source text as possible.” (Culy 2003)
  - References must be similar, to give meaningful results, but Existing translations are too divergent to be useful
    - Tom Sawyer, Bible translations (Culy 2003)
- This problem does not go away with multiple references
  - (Some papers report it helps, some report it does not)

Test translations?

Scores seem to be sensitive to kind of system taken for test translation

- Are not able to distinguish between human translations **("The human translations that scored poorly were generally freer translations" Culey 2003)**
- Some system types are liked more by scoring procedures than others
  - BLEU like SMT more than RMT (Och 2005)

Human Evaluation?

- Evaluation scores of humans depend on
  - Language skills
    - (monolingual vs. bilingual evaluators)
  - Domain expertise
  - Kind of material to be evaluated
    - Some evaluators seem to prefer RMT errors over SMT errors
  - Way of material presentation
    - Adequancy judgements influenced by prior fluency evaluation (Sourcin)
    - Reference translation presentation influences human scoring **(“human raters appeared very reluctant to give a perfect 4.0 rating to translations that did not exactly match the reference” (Coughlin)).**
    - Availability of context vs. of only single sentences (Coughlin)
    - Evaluation setup (speed / tedious task)
  - The nature of the task is adverse

    “the superior linguistic skills of human raters are not exploited by MT evaluation tasks that involve quickly comparing a machine-tanslated sentence to a human-translated reference … they behave like expensive, slow versions of BLEU” (Coughlin)

Possibility 2: Content of Metrics

- Result 1: Divergence in ocngruency of hman / automatic metrics may be due to some paramters of the settings
  - Would need to be further investigated

- Option 2: Divergence is due to intrinsic problems of the applied metrics themselves:

  - Objective: Rank the most similar document best
  - Main heuristics:
    - Word error rate (Same vocabulary as in reference?)
      - Approximation of correct lexical selection
    - N-gram calculation (Same *sequence* of words as in reference?)
      - Approximation of syntactic ordering
  - Different ways of combining the two

Are these heuristics valid / intuitive?

Word ‘errors’

- Are different words really errors? Some are, some are not.
- ***Variance*** of concepts is a prominent means of good translation. This is not covered by WER

- WER measures *deviation* but neglects its ***quality***

Take the floppy out of the drive	*Take the disquette out of the drive*
	**Take the elefant out of the drive*
Hans ging nach Hause zurück	*John went back home*
	John returned to his home
	**John went from home*

  It is not clear if a deviation improves or deteriorates results

- This holds also for technical terms:
  - In MT evaluation: *Fidelity? Adequacy? Accuracy?*
  - linguatec: term bank of 2.046 mio bilingual entries, has 2.03 translations of a term on average

Examples 1

Kreuzkopf	*cross head*	[bigram]
	piston crosshead	
	tie-bar	
	cross high filler spot	
	crosstail	
	cross heads	
	crosshead	[unigram]
Kreuzkopfmotor	crosshead engine	
	cross high filler spot engine	
Kreuzlenker	cross handlebar	
	transversely jointed linkage	
	cross-shaped link	
	cross steering wheel	
	cross suspension link	
Kreuzpflaster	cross plaster	
	cross-ply patch	
	cross-ply boot	
Kreuzriffelung	cross checker	
	cross-wise corrugation	

Examples 2

Kreuzschlitzkopf	intended cross head phillips recessed head cross recessed head *phillips* head crosshead head	
Kreuzspeichen	*crossing* spokes *crossed* spokes cross wire spokes	[RED ☺ BLEU ☹]
kreuzungslos	intersection-free *without intersections* **intersection batch** nonintersecting grade-separated	 [different POS] [?]
kreuzweise wechselnd	crosswise changing	
Kreuzwickel	*cheese* cross-wound bobbin cross winding cross compress	

n-gram contexts

- N-grams are intended to prefer phrases / constituents
- This is language dependent
  - Languages differ in freedom of constituent / word order
- It neglects the fact that sequence of words carries meaning

the letter opened yesterday =>	[participle reading]
yesterday the letter opened	[different meaning]
letter opened yesterday the	[illformed]

=> Both word error rate and n-gram similarity are not terribly intuitive heuristics for meaning preservation

Examples 1

DE: Im R/3-System unterscheiden wir zwischen verschiedenen Typen von ABAP-Programmen
T1: in the R /3 System , there are various types of ABAP programs
T2: We distinguish between different types of ABAP programs in the R/3 system.

DE: tragen Sie sich bitte als Pilot ein , und ordnen Sie sich einem Flug zu.
HUM: enter yourself as pilot and assign yourself to a flight .
MT1: enter this as a pilot , and assign a flight .
MT2: Please put down your name as a pilot and assign yourself to a flight.

DE: Mit dieser Methode ermitteln Sie den ersten Kindknoten eines bestimmten Knotens.
HUM: use this method to find out the first child node of a given node .
MT1: use this method to find out the first child node of a given node .
MT2: With this method you find out the first child node of a certain node.

Examples 2

DE: Die effizientesten Schreiboperationen werden am Checkpoint durchgeführt (“Chunk Writes”)
HUM: the most efficient writes are done at checkpoint (“chunk writes”)
MT1: the writes to the checkpoint to be performed (“chunk Writes”)
MT2: The most efficient write operations are executed at the checkpoint (“Chunk Writes”)

DE: Es gibt zwei unterschiedliche Möglichkeiten, aus ABAP-Programmen auf Datenbanken zuzugreifen:
HUM: there are two ways to access the database from an ABAP program:
MT1: there are two different ways from ABAP programs access databases:
MT2: There are two different possibilities of accessing databases from ABAP programs:

DE: Von ihrem Arbeitsplatz aus sollen Ihre Mitarbeiter auch vereinfachte Transaktionen ausführen können .
HUM: your colleagues should also be able to execute simplified transactions from their workplaces .
MT1: from their workplace should your employee also simplified transactions .
MT2: Your employees also shall be able to execute simplified transactions from their workstation

Examples 3

DE: Für die Verknüpfung mit Organisationsobjekten müssen Sie kein HR im Einsatz haben.
HUM: you do not have to have HR for the relationships with organizational objects .
MT1: for the link with organizational objects must use in HR you do not have .
MT2: You don't have to have any HR for the bonding with organization objects in use.

DE: Die Parameter der Datenbanksystemprüfung können Sie in der R/3- Tabelle DBCHECKORA konfigurieren.
HUM: configure the database system check parameters in the R /3 table DBCHECKORA .
MT1: the parameters of the database system check in R /3 configure table DBCHECKORA .
MT2: The parameters of the database system test can configure you in the R/3 table DBCHECKORA.

Metrics: Word error rate

- Given such data, which metrics are proposed to discriminate? ->
- Calculate distance of documents on a single word basis: Compare the words of test and reference text:
- **Word Error Rate**, calculated as edit distance (Niessen)

  Edit distance also used in RED (Akiba et al.)
- Unigram analysis, and f-measure (Turian et al. 2003)

  “There is no point in comparing MT systems on word order when all MT systems are equally disfluent.”

Word Context

- Next observation: 5 words in row better than 5 words garbled: Not reflected by WER.
- Possible Reactions:
  - Take this into account
    - Identify and prefer word blocks (Leusch)
    - Punish misplaced tokens (Marrafa)

    BLEU / NIST: use ngram information in addition to WER
    - But: relative weight of unigrams vs. multigrams?
      - NIST: 97% of score determined by unigrams + bigrams
      - BLEU: Some 3-4-grams tolerate > 30% wrong unigrams (Zhang)
    - Other properties:
      - Prefer SMT output over RMT output (Och)
      - Loose expressive power for very good and very bad tanslations (Och)
        - » Humans have bad BLEU scores
  - Ignore it
    - Create a position independent error rate (PER) (Leusch)

Weighted Ngrams, f-measure

- Next observation: Words are not equally important:
  - Calculate **information load** of words, based on tf:idf score
    WNM: weighted ngram measure
    - correct *of the* should not outperform false *Nicaragua*
    - (Babych, Rajman, Hartley)

- Current tendency:
  Use the Recall / Precision paradigm for MT Evaluation
  - Hartley, Rajman, et al., Turian, Melamed
  - Recall / precision sometimes are used in a metaphoric sense …

(Still all of this is based on counting ‘word errors’)

(Increasing similarity instead of quality)

- Once the focus is on recall and precision:
  There are means to improve retrieval performance:
  - Stemming, case normalisation etc. gives better retrieval results
  - Replacing steps by semantic tags or POS levels (Turian, Akiba)
- However, evaluation started from a *text*
  - With full forms, cased, …

(Optimising similarity is not always optimising translation quality …)

Results

1. Metrics based on document similarity are questionable:
   - They give **controlled metrics in an uncontrolled setting**
     - Too many uncontrolled parameters
     - Reference translations as uncontrollable as human evaluation judgements
   - They are **based on non-intuitive heuristics**
     - Word errors, n-grams
   - They have **effects which are counterintuitive**
     - Better translations score worse
     - Human translations have bad scores
   - We dont know when the results are **valid** / you cannot trust them
     - You do not know *when* they are in line with human judgement
   - They are **not informative**
     - What does a BLEU score of 3.6 mean?
   - They do not measure : Fluency, i.e.
     Produce a **correct target sentence** (a basic translation requirement)
     - All human schemes have a notion of ‘correctness’

Results

2. Metrics based on common SL-TL features seem to be better indicators for translation quality.
   - Has not been systematically tried yet

3. The automatic metrics can be good for:

   - Offering a ***ranking*** of different systems

   - Measuring overall progress of the ***same system*** in a certain period of its lifetime

Results

4. They are not good for:
   - R&D on system improvement
     "Automatic metrics are not designed to provide direction to R&D" (Miller)
   - They give no guide what to do next

   They cannot replace in-depht / finegrained evaluations:
   - Dialogue acts, predicate-argument structures (Belvin)
   - Lexical gaps, tag and segment errors, unknowns, names (Correa)
   - Syntax, Morphology, Lexicon (Miller)
   - NPs, VPs, Lexical analysis (Mustafa)

   To find out that 30% of your errors are unknown words you don't need a BLEU score – and BLEU would not tell you

Remember …

We wanted a cheap, objective, replicable, fast, automatic solution

- Cheap? "building test suites is expensive" (Zhang)
  meant for reference translation production!
- Objective? Settings have too many unknown parameters
  All efforts include human judgement
- Replicable? Redo for every corpus, language, setting, …
- Automatic? OK
- Fast? OK

BUT: Use of automatic means is necessary

linguatec
Sprachtechnologien

Thank you for your attention

g.thurmair@linguatec.de