

---

# Workpackage 3

## Interactive Editing

Philipp Koehn, Jesús González-Rubio

November 25, 2013



# Overview

- Objective: New methods to assist the editing of translations
- Tasks in Year 2
  - Task 3.1: Sentence-level Estimate of Post-editing Work Effort (completed)
  - Task 3.2: Word-level Confidence Measures (completed)
  - Task 3.5: Display Multiple Translation Options (ongoing into year 3)
  - Task 3.6: Authoring Assistance (ongoing into year 3)

## Shared Task at WMT 2013

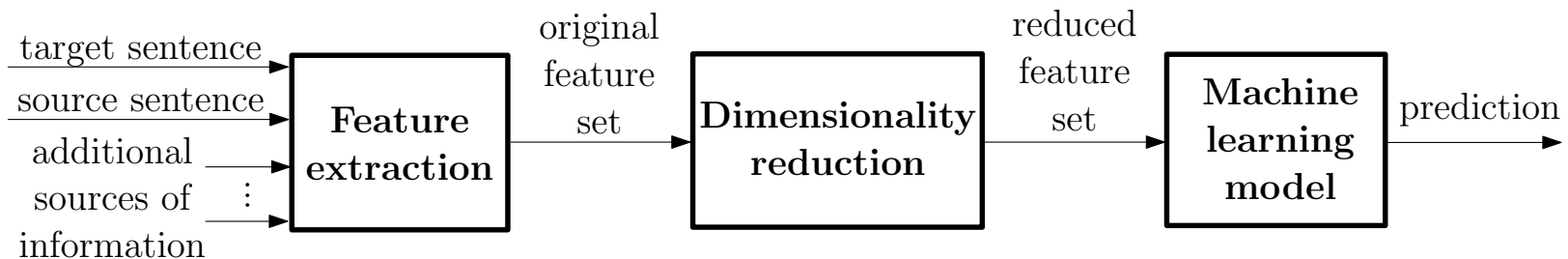
- Much interest on confidence measures (*quality estimation*) in research community
- ⇒ Organization of shared task  
(collaboration with MATECAT, MOSESCORE, QTLAUNCHPAD)
- WMT workshop at ACL conference, Sofia, Bulgaria, August 2013
- Tasks
  - prediction of usefulness of sentence translations
  - prediction of post-editing time
  - word level confidence estimation
- Tasks used data from first CASMACAT field trial as test

## Task 3.1

# Sentence-Level Estimate of Post-Editing Work Effort

## Task 3.1: Sentence-Level Estimate of Post-Editing Work Effort

- Schedule: months 1–24
- Extension of the work developed during the first year
  - Regression problem, prediction of quality scores from a set of features
- Focus: efficient management of huge sets of collinear and ambiguous features
- Proposal: A two-step training methodology



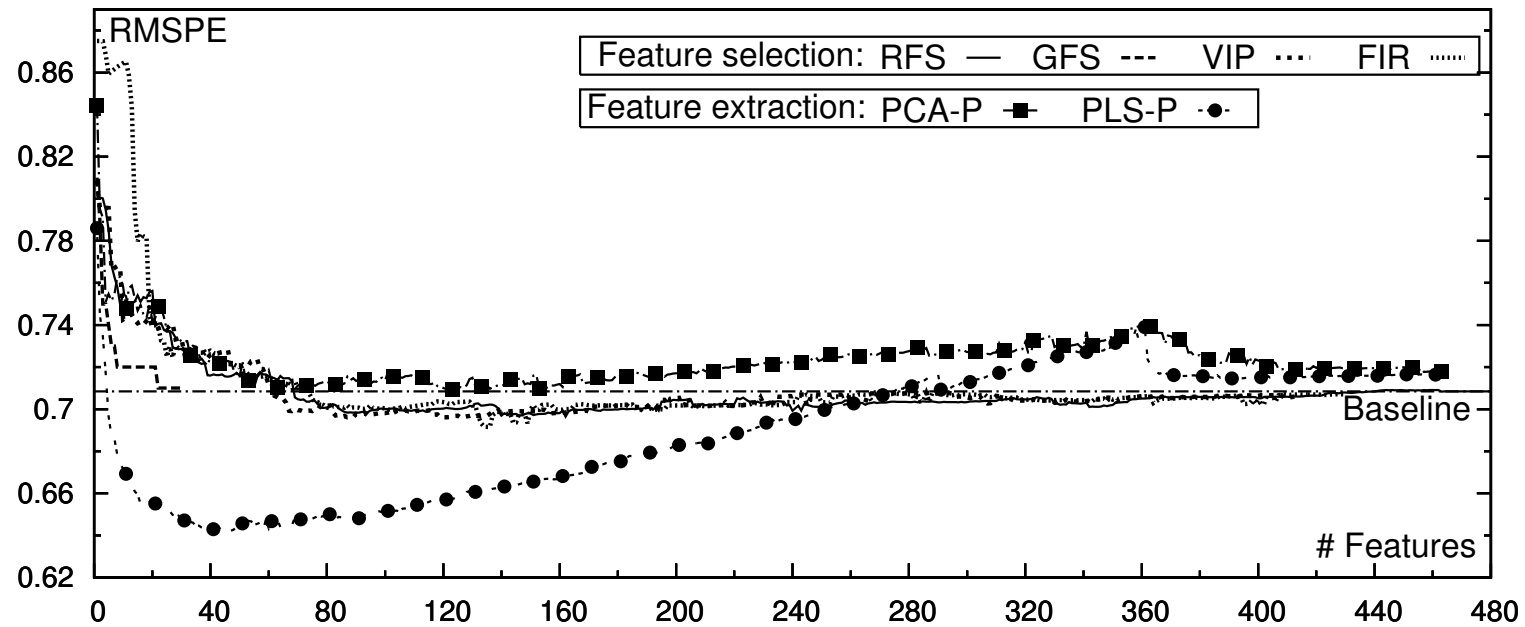
## Development

- Dimensionality reduction (DR) based on *partial least squares regression* (PLSR)
- Widely-used PCA reduces dimensionality taking into account only the features
  - Reduced feature sets contain almost not redundancy...
  - but, these are not necessarily the best features to perform the prediction
- Alternatively, PLSR does take into account the values to be predicted
  - As PCA, reduced feature sets contain almost no redundancy...
  - and, the new features explain most of the variability in the values to be predicted

## Contributions

- Two new DR methods based on PLSR:
  - PLSR projection:** projects the original features into a new space (similar to PCA)
  - Variable importance in projection:** selects a subset of the original features
- Exhaustive comparison against different widely-used DR methods
  - Measure of interest: prediction accuracy when using each DR method
- Study the influence of each DR method on the accuracy of different prediction models

# Results



- PLSR projection (PLS-P) outperformed all other tested approaches
- Huge reduction in the number of features used to perform the prediction
- Similar conclusions were obtained for all tested prediction models



## Conclusions

- Projection-based DR methods usually outperformed feature selection methods
  - The proposed PLS projection outperformed widespread PCA projection
- A combination of PLSR-P and a SVM provided the best performance
  - Better prediction accuracy than models built with all the original features
- Time efficiency is a complimentary advantage of the proposed approach
  - Adequate approach to be deployed in scenarios with temporal restrictions
- The ideas explored here influenced the implementation of Active Learning in WP4

## Publications

- Jesús González-Rubio, José R. Navarro-Cerdan, Francisco Casacuberta. Partial Least Squares for Word Confidence Estimation in Machine Translation. *6th Iberian Conference on Pattern Recognition and Image Analysis, (IbPRIA) LNCS 7887*, 2013. pp. 500-508. Springer.
- Jesús González-Rubio, José R. Navarro-Cerdan, Francisco Casacuberta. Empirical Study of a Two-Step Approach to Estimate Translation Quality. *Proceedings of the International Workshop on Spoken Language Translation (IWSLT)*, 2013.

## Task 3.2

# Word-Level Confidence Measures

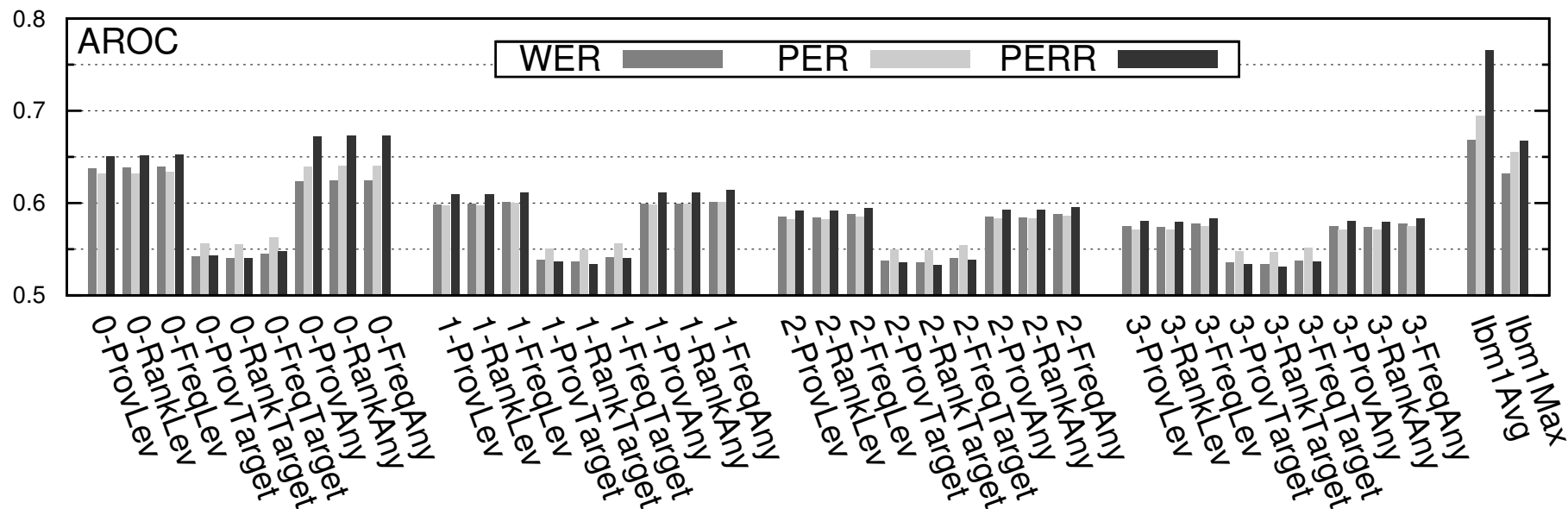
## Task 3.2: Word-Level Confidence Measures

- Schedule: months 1–24
- Confidence estimation is addressed as a two-class classification task
- Focus: efficient management of large sets of noisy features
- Proposal: classifier based on *partial least squares discriminant analysis* (PLS-DA)
- Empirical study of the influence of context in classification accuracy

## Development

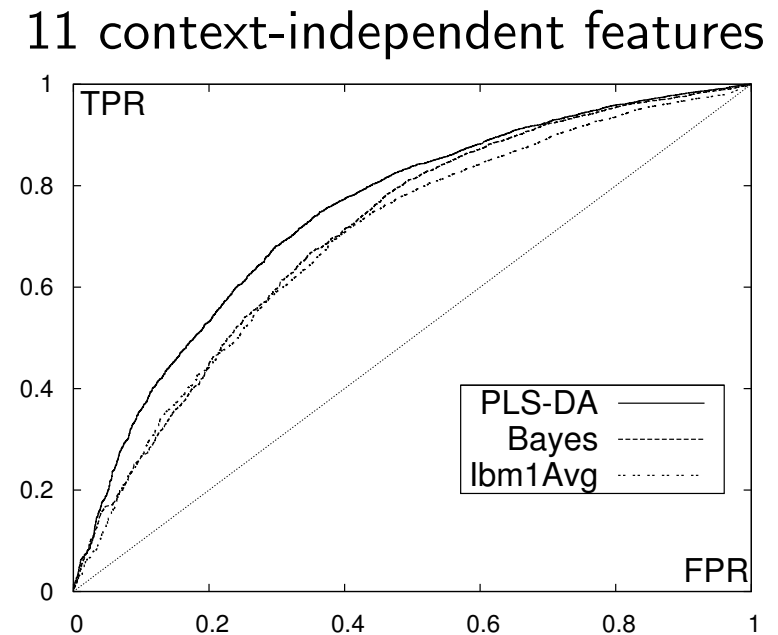
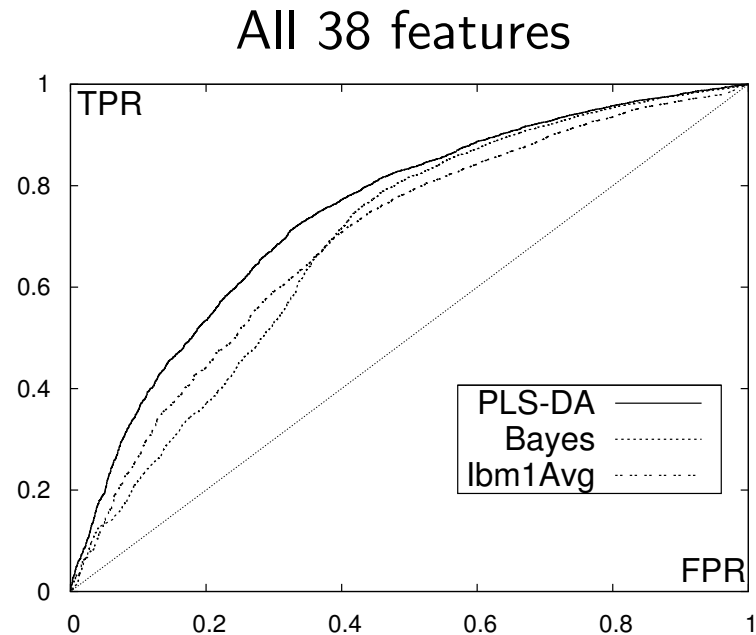
- Study of new features describing context
  - Generalization of the features based on posterior probabilities
  - Context is included as an additional posterior dependency
  - Drawback: redundancy
- PLS-DA classifier
  - Performs an implicit dimensionality reduction based on PLSR projection
  - Efficient management of large sets of noisy and redundant features
- Empirical study of each individual feature
- Comparison against the state-of-the-art in word-level quality estimation

## Results of the Individual Features



- Prediction accuracy degraded as context size increased
- Features based on Model 1 lexicon outperformed posterior features

## Results of the PLS-DA Classifier



- PLS-DA outperformed the previously used Naïve Bayes classifier
- Robust when multiple redundant features (different context sizes) were used

## Conclusions

- The best individual feature is based on a Model 1 lexicon
  - This quality estimator is thus chosen to implement Active Interaction in WP2
- Good individual performance of context-aware features, however:
  - Larger contexts reduce the individual accuracy of the features
- PLS-DA outperformed previous models
  - Better performance in all test conditions
  - Empirical robustness in the presence of redundant and noisy features
  - Scalable due to the implicit dimensionality reduction



## Publication

- Jesús González-Rubio, José R. Navarro-Cerdan, Francisco Casacuberta. Dimensionality reduction methods for machine translation quality estimation. *Machine Translation*, 2013. Vol. 27 (3), pp. 281-301.

## Task 3.5

# Display of Multiple Translation Options

## Translation Option Array

er	hat	seit	Monaten	geplant	,	im	März	einen	Vortrag	...
he has		for months		the plan		in March		a lecture		...
it has		for months now		planned	,	in	March	a presentation		...
he was		for several months		planned to		in the March		a speech		...
he has made		since	months	the pipeline		in March of		a statement		...
he did		for many months		scheduled		the March		a general		...

- Main work in Year 3
- Focus
  - diversity of displayed options
  - user studies on how much to display
  - user interface issues

# Translation Options in Context

Speaking in Latin to a small gathering of cardinals at the Vatican on Monday morning, Benedict

said

stre

world

The s

A shy

Bene

An often divisive figure, he spent much of his papacy in the shadow of his beloved predecessor.

**gathering**

ds and fish when not out **gathering** edible plants. re, Fische und Vögel und **sammeln** essbare Pflanzen.

mpact assessment and by **gathering** information. hmen und Informationen **sammeln** .

ace is already capable of **gathering** information on the most i europäischen Sender zu **sammeln** , aufzubereiten und zur \

**Treffen**

inform better about our **gathering** "Don Bosco weltweit" we ind Jugendgruppen unser **Treffen** "Don Bosco weltweit" be

ember 2nd 's pro-nuclear **gathering** in Lyon - EFN officially n - Pro-nukleares **Treffen** in Lyon am 2. Septembe

ladies and gentlemen, a **gathering** involving the Health Con und Kollegen! bei einem **Treffen** vor einigen Tagen in Brü

ment in that regard at a **gathering** of displaced persons. che Erklärung auf einem **Treffen** von Vertriebenen abgege

**Erfassung**

systems - requirements **gathering** , analysis and developm ysteme - Anforderungen **Erfassung** , Analyse und Entwicklur

is and prevention and on **gathering** and processing epidemio e, die Prävention und die **Erfassung** und Verarbeitung von Da

ased but comprehensive **gathering** of data on all chemicals, reisen aber umfassenden **Erfassung** von Informationen zu sä

I realise that the **gathering** of statistics needs to be ich erkenne an, dass die **Erfassung** von statistischen Daten \

**Erhebung**

as far as the **gathering** , processing and use of ; bei der **Erhebung** , Verarbeitung und Nutz

er States, as well as the **gathering** and analysis of reliable s zu; das gilt auch für die **Erhebung** und Auswertung verlässli

## Main Concept

- Find source word/phrase in parallel corpus
- Display possible translations
  - ranked by frequency
  - in sentence context
  - aligned phrase highlighted

## Technical Aspects

- Uses parallel corpus, just as for training of machine translation systems
- Standard automatic word alignment techniques
- Data stored in suffix array

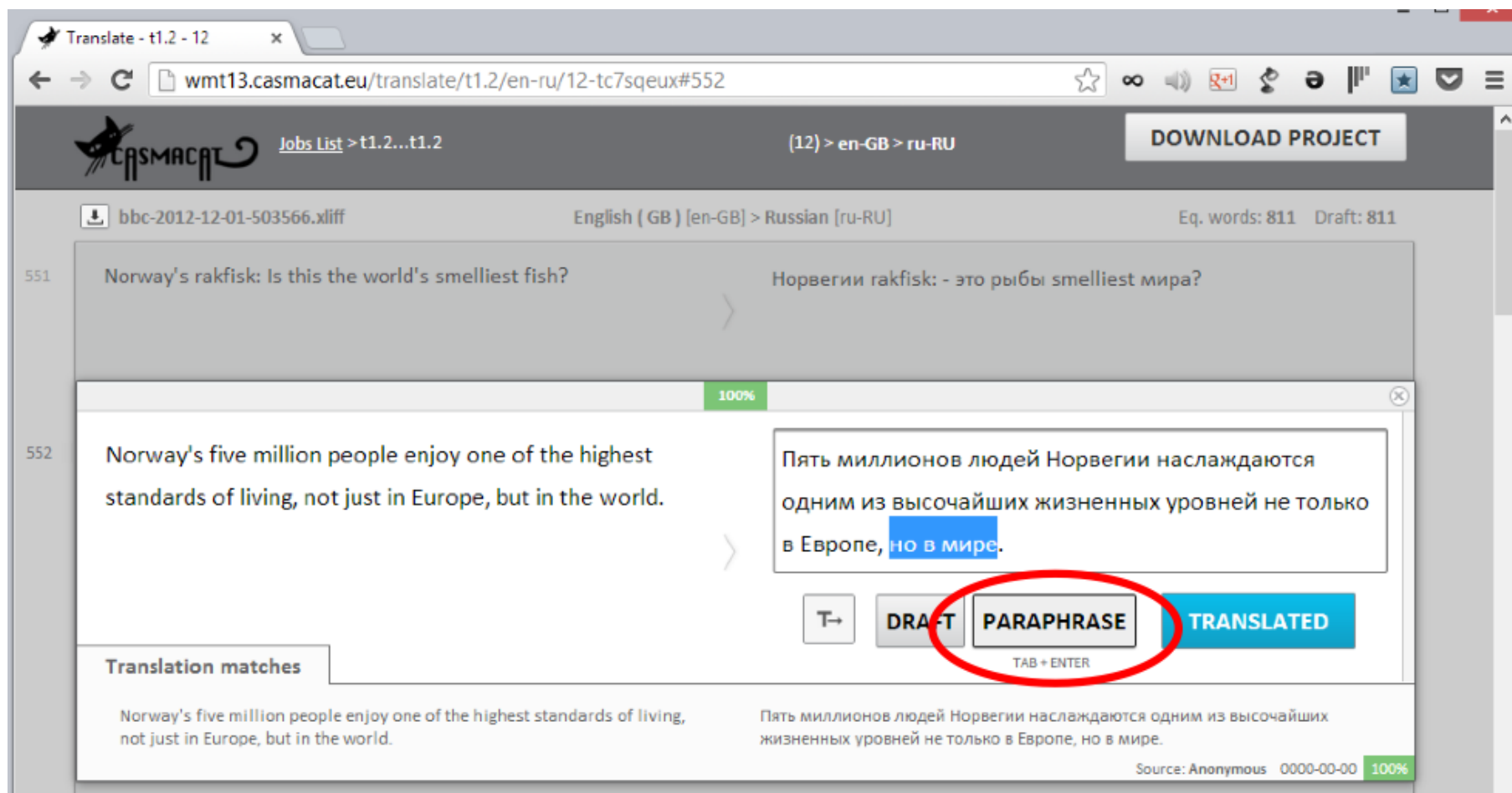
⇒ very fast retrieval

## Task 3.6

# Authoring Assistance

# Display of Alternative Translations

## Request



The screenshot shows a web browser window with the URL `wmt13.casmacat.eu/translate/t1.2/en-ru/12-tc7squeux#552`. The interface displays a translation project for the file `bbc-2012-12-01-503566.xliff`, with the source language set to English (GB) [en-GB] and the target language to Russian [ru-RU]. The project statistics show 811 equivalent words and 811 draft words.

Line 551 shows the source text: "Norway's rakfisk: Is this the world's smelliest fish?" and the target translation: "Норвегии rakfisk: - это рыбы smelliest мира?".

Line 552 shows the source text: "Norway's five million people enjoy one of the highest standards of living, not just in Europe, but in the world." The target translation is: "Пять миллионов людей Норвегии наслаждаются одним из высочайших жизненных уровней не только в Европе, но в мире." The word "но" is highlighted in blue.

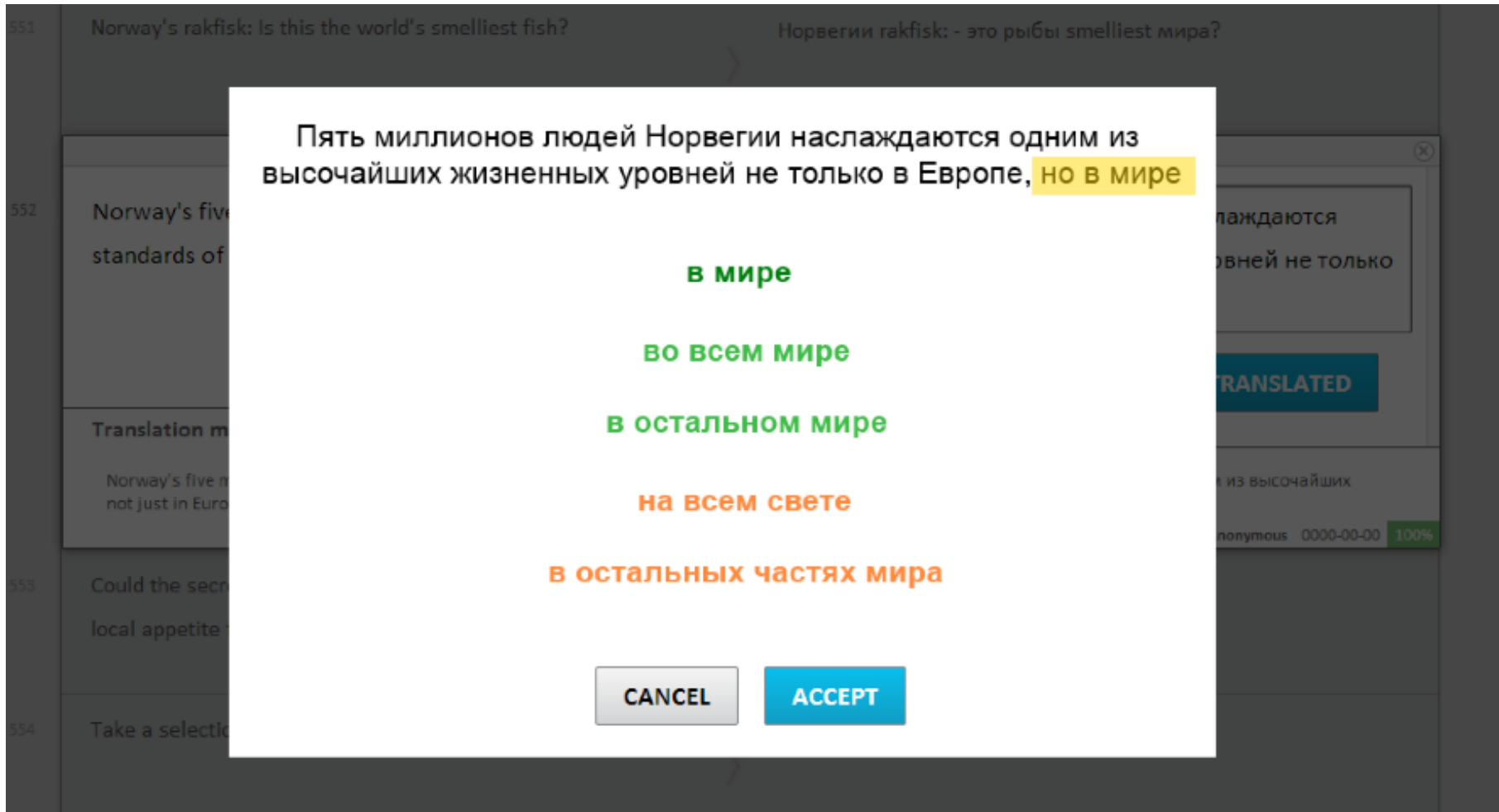
Below the translation, there are buttons for "T→", "DRAFT", "PARAPHRASE", and "TRANSLATED". The "PARAPHRASE" button is circled in red. Below these buttons, the text "TAB + ENTER" is visible.

A "Translation matches" section is also present, showing the source and target text side-by-side for comparison. At the bottom right, the source is identified as "Anonymous" with a 100% match rate.



# Display of Alternative Translations

## Display



551 Norway's rakfisk: Is this the world's smelliest fish? Норвегии rakfisk: - это рыбы smelliest мира?

552 Norway's five standards of life are not just in Europe, but in the world.

553 Could the secret of local appetite be found in the world's most delicious food?

554 Take a selection of the world's most delicious food.

Пять миллионов людей Норвегии наслаждаются одним из высочайших жизненных уровней не только в Европе, **но в мире**

- в мире**
- во всем мире**
- в остальном мире**
- на всем свете**
- в остальных частях мира**

CANCEL ACCEPT

TRANSLATED

полностью 0000-00-00 100%

## Relation to Paraphrasing

- Automatic paraphrasing established research topic
- Our application:
  - driven by source
  - use of search graph
  - considers sentence context
  - ranking and diversity important
  - real time performance

## Generation

- User marks out part of translation (target span)
- Target span is mapped to source span
- Search graph is consulted for alternative translations for source span
- Additional translations generated by combining translation options

⇒ Initial list of translations

- Note: could also use monolingual paraphrasing resources

# Components

- Partial filters: remove some translation options
- Scorers: score translations
- Filters: remove some translations from list
- Sorters: rank list

## Partial Filters

Remove some bad phrase translations

- PTPF: Punctuation Partial Filter
- SDPF: String Distance Partial Filter  
remove phrase translations that are too similar to others
- FWPF: FunctionWord Partial Filter  
remove phrase translations if they have additional function words

# Scorers

Scores each translation

- BFSF: Best Forward Score Function  
compares alternate translations against best path
- SDSF: Score Difference Score Function  
considers direct and indirect conditional probability
- LMSF: Language Model Score Function

## Filters

Remove some translations for full span  
(not just partial phrase translations)

- Span versions of phrase translation filters
- SBRF: Score Based Filter  
remove phrase translations with bad overall score

# Sorters

Sorts entire final list of translations

- LMBS: Language Model Based Sorter  
uses full sentence language model score
- CBDS: Cluster Based Diversity Sorter  
first clusters translations by similarity  
then picks best translation from each cluster  
(k-means clustering)



## Automatic Evaluation

- Motivation
  - alternative translations should fix translation errors
  - create bad translations by back-translation
- Process
  - Train machine translation system for both directions
  - Translate test set target → source → target\*
  - Spot differences between target and target\*
  - Use span in target\* as “marked by user”, span in target as correct
- Experimental setting
  - WMT 2013 news translation task
  - English–Russian
  - Display 5 alternate translations (one has to be correct)
  - 2139 test cases from 1000 sentences

## Example

- Translate

*Unlike in Canada , **the American states** are responsible for the organisation of federal elections.*

- Into

**В отличие от Канады, американские штаты ответственны за организацию федеральных выборов в соединенных штатах .**

- Back into English

*Unlike in Canada , **US states** are responsible for the organization of federal elections.*

## Results

Method	Partial Filters	Score Function	Filters	Sorters	Match
1	PTPF	SDSF	-	-	135/2139
2	PTPF, FWPF, SDPF	SDSF	-	-	161/2139
3	PTPF, FWPF, SDPF	BFSF	-	-	159/2139
4	PTPF, FWPF, SDPF	LMSF	-	-	211/2139
5	PTPF, FWPF, SDPF	SDSF, BFSF	-	-	161/2139
6	PTPF, FWPF, SDPF	SDSF, LMSF, BFSF	-	-	218/2139
7	PTPF, FWPF, SDPF	SDSF, LMSF, BFSF	PTRF, FNRF	LMBS	574/2139
8	PTPF, FWPF, SDPF	SDSF, LMSF, BFSF	PTRF, FNRF	CDBS	493/2139
9	PTPF, FWPF, SDPF	SDSF, LMSF, BFSF	SBRF	LMBS	661/2139
10	PTPF, FWPF, SDPF	SDSF, LMSF, BFSF	SBRF	LMBS,CDBS	691/2139

## Manual Evaluation

- Web based interactive evaluation tool
- Same setup as automatic evaluation
  - shows target span
  - 5 selectable paraphrases
  - user accepts one → correct
- Limit to three approaches
- Limit to 50 test cases

## Results

- Four users (U1–U4)
- Number of instances where one translation is correct

Method	U1	U2	U3	U4	average score
1	8	6	9	6	6/50
7	15	17	12	10	13/50
10	24	20	26	29	26/50

**Thank You**

questions?