# WMT 10 Shared Tasks:
# Translation Task
# System Combination Task

Chris Callison-Burch, Philipp Koehn, Christof Monz, Omar Zaidan

15 July 2010

# Translation Task

- Open benchmark for machine translation

- Every year since 2005, we ...
  - post training data on a web site
  - prepare a test set
  - given participants 5 days to translate the test set
  - score the results

- 8 language pairs (Czech, German, French, Spanish ↔ English)

- Sponsored by the EuroMatrixPlus project (EU FP7)

# Machine Translation Marathon

- If you have a new graduate student ...
  → send her to a 1-week intensive hands-on SMT course

- If you have developed a open source tool for MT
  → submit a paper to the open source convention (deadline August 1)

- If you want to get practical experience in MT code
  → join the one-week hack fst

- All this at the 5th MT Marathon
  - Le Mans, France, September 13-18, 2010
  - http://lium3.univ-lemans.fr/mtmarathon2010/

# What’s New?

- Professionally translated test set (by EuroMatrixPlus partner CEET)
- More data – for some language pairs vastly more data
- Added manual evaluation with Mechanical Turk
- Metrics evaluation handled by NIST (will be presented tomorrow)

# Participants

- 29 Institutions
  - Europe: 21
  - North America: 7
  - Asia: 1
- 33 groups
- 153 submitted system translations, also included
  - two popular online translation systems
  - rule-based systems for English–Czech

# Training Corpora

- Updated Europarl (50MW) and News Commentary (2MW) releases
- Updated monolingual news corpora (100-1100MW)
- Much larger 120MW Czech-English corpus (by Ondrej Bojar)
- New 200MW UN corpus for Spanish–English and French–English (by DFKI)

# Test Set

- News stories

- Sources taken from 5 different languages

  **Czech:** iDNES.cz (5), iHNed.cz (1), Lidovky (16)
  **French:** Les Echos (25)
  **Spanish:** El Mundo (20), ABC.es (4), Cinco Dias (11)
  **English:** BBC (5), Economist (2), Washington Post (12), Times of London (3)
  **German:** Frankfurter Rundschau (11), Spiegel (4)

- Translated across all 5 languages (multi-lingual sentence aligned corpus)

# Manual Evaluation

- **Sentence Ranking**: Which systems are better?
  Rank translations from Best to Worst relative to the other choices (ties are allowed).

- **Sentence Correction**: How understandable are the translations?
  - stage 1: Editing the translation (w/o source and reference)
    Correct the translation displayed, making it as fluent as possible. If no corrections are needed, select "No corrections needed." If you cannot understand the sentence well enough to correct it, select "Unable to correct."
  - stage 2: Assessing the correctness (with source and reference)
    Indicate whether the edited translations represent fully fluent and meaning-equivalent alternatives to the reference sentence. The reference is shown with context, the actual sentence is **bold**.

# Mechanical Turk

- Platform to crowd-source online tasks (very cheap: $.05 for 3 rankings)
- Main problem: quality control
- Requirements for workers
  - existing approval rating of at least 85
  - must have at least performed 5 task
  - resides in a country where target language is spoken

# Evaluations Collected

- Goal: 600 ranking sets per language pair, each posted redundantly 5 times

- Actual:

| | en-de | en-es | en-fr | en-cz | de-en | es-en | fr-en | cz-en |
|---|---|---|---|---|---|---|---|---|
| Location | DE | ES/MX | FR | CZ | US | US | US | US |
| Completed 1 time | 37% | 38% | 29% | 19% | 3.5% | 1.5% | 14% | 2.0% |
| Completed 2 times | 18% | 14% | 12% | 1.5% | 6.0% | 5.5% | 19% | 4.5% |
| Completed 3 times | 2.5% | 4.5% | 0.5% | 0.0% | 8.5% | 11% | 20% | 10% |
| Completed 4 times | 1.5% | 0.5% | 0.5% | 0.0% | 22% | 19% | 23% | 17% |
| Completed 5 times | 0.0% | 0.5% | 0.0% | 0.0% | 60% | 63% | 22% | 67% |
| Completed $\geq$ once | 59% | 57% | 42% | 21% | 100% | 99% | 96% | 100% |
| **Label count**<br>**(% of expert data)** | **2,583**<br>**(38%)** | **2,488**<br>**(96%)** | **1,578**<br>**(40%)** | **627**<br>**(9%)** | **12,570**<br>**(241%)** | **12,870**<br>**(228%)** | **9,197**<br>**(222%)** | **13,169**<br>**(490%)** |

# Intra and Inter-Annotator Agreement

**Inter-annotator agreement**

| | $P(A)$ | Kappa | Kappa experts |
|---|---|---|---|
| With references | 0.466 | 0.198 | 0.487 |
| Without references | 0.441 | 0.161 | 0.439 |

**Intra-annotator agreement**

| | $P(A)$ | Kappa | Kappa experts |
|---|---|---|---|
| With references | 0.539 | 0.309 | 0.633 |
| Without references | 0.538 | 0.307 | 0.601 |

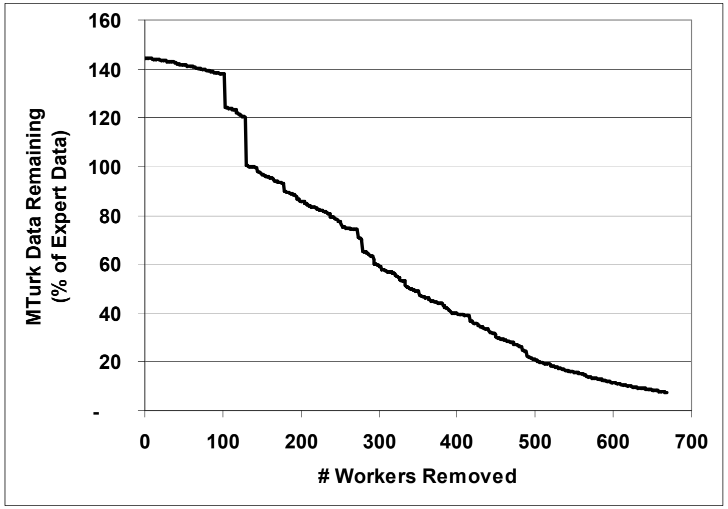
# Detecting Bad Workers

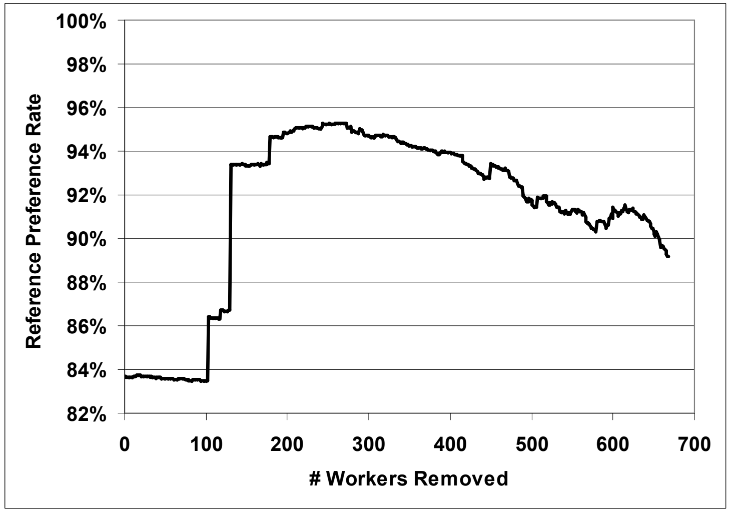
- Indicators
  - low *reference preference rate* $(RPR)$: prefer MT output often over references
  - low agreement with experts

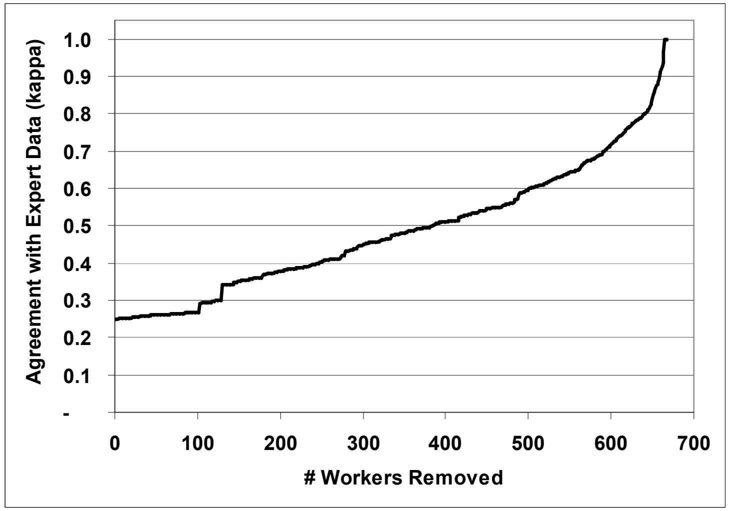
⇒ Filter out the bad workers

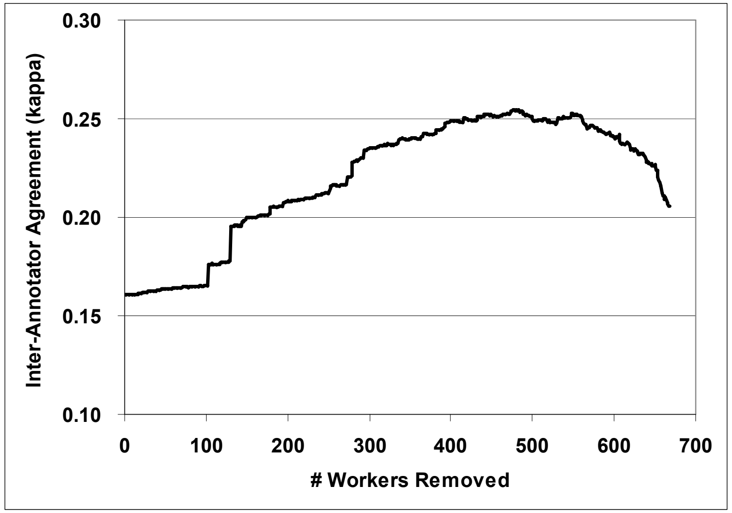
- Very few workers have to removed for better quality (two worst offenders responsible for most damage)

# Removing Bad Workers









# Spearman Rank Coefficients

Comparing MTurk rankings with Expert rankings

| | Label count | Unfiltered | Voting | $K_{exp}$ filtered | $RPR$ filtered | Weighted by $K_{exp}$ | Weighted by $K(RPR)$ |
|---|---|---|---|---|---|---|---|
| en-de | 2,583 | 0.862 | 0.779 | 0.818 | 0.862 | **0.868** | 0.862 |
| en-es | 2,488 | 0.759 | 0.785 | 0.797 | 0.797 | 0.768 | **0.806** |
| en-fr | 1,578 | 0.826 | **0.840** | 0.791 | 0.814 | 0.802 | 0.814 |
| en-cz | 627 | 0.833 | 0.818 | 0.354 | 0.833 | **0.851** | 0.828 |
| de-en | 12,570 | 0.914 | 0.925 | 0.920 | 0.931 | **0.933** | 0.926 |
| es-en | 12,870 | 0.934 | 0.969 | 0.965 | **0.987** | 0.978 | **0.987** |
| fr-en | 9,197 | 0.880 | 0.865 | **0.920** | 0.919 | 0.907 | 0.917 |
| cz-en | 13,169 | 0.951 | 0.909 | **0.965** | 0.944 | 0.930 | 0.944 |

# Results

- Conditions
  - systems may only use the provided data (constraint)
  - systems may use additional data (unconstraint)
  - systems may use the LDC Gigaword corpus (GW)

- Ranking
  - systems are ranked by how often they were ranked $\geq$ any other system.
  - ties are broken by direct comparison.
  - ● indicates a **win** in the category, meaning that no other system is statistically significantly better at p-level$\leq$0.1 in pairwise comparison.
  - ⋆ indicates a **constraint win**, no other constraint system is statistically better.

- For all pairwise comparisons between systems, please check the paper.

# Pairwise Comparison

| | REF | AALTO | CMU | CU-BOJAR | CU-ZEMAN | ONLINEA | ONLINEB | UEDIN | BBN-C | CMU-HEA-C | JHU-C | RWTH-C | UPV-C |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| REF | – | .03‡ | .02‡ | .03‡ | .01‡ | .03‡ | .02‡ | .05‡ | .02‡ | .06‡ | .03‡ | .05‡ | .03‡ |
| AALTO | **.93**‡ | – | **.54**‡ | **.54**‡ | .23‡ | .36 | **.58**‡ | **.56**‡ | **.65**‡ | **.69**‡ | **.64**‡ | **.67**‡ | **.62**‡ |
| CMU | **.94**‡ | .30‡ | – | **.47** | .14‡ | .22‡ | **.52**‡ | **.41** | **.50**‡ | **.57**‡ | **.45**† | **.44** | .38 |
| CU-BOJAR | **.94**‡ | .26‡ | .38 | – | .10‡ | .22‡ | **.61**‡ | **.47**† | **.46** | **.55**‡ | **.42** | **.49**‡ | **.44** |
| CU-ZEMAN | **.98**‡ | **.58**‡ | **.73**‡ | **.77**‡ | – | **.55**‡ | **.79**‡ | **.71**‡ | **.84**‡ | **.80**‡ | **.77**‡ | **.79**‡ | **.75**‡ |
| ONLINEA | **.94**‡ | **.41** | **.61**‡ | **.57**‡ | .23‡ | – | **.68**‡ | **.63**‡ | **.71**‡ | **.71**‡ | **.63**‡ | **.54**‡ | **.61**‡ |
| ONLINEB | **.93**‡ | .30‡ | .31‡ | .26‡ | .10‡ | .17‡ | – | .32† | .35 | .31 | .22‡ | .29⋆ | .38 |
| UEDIN | **.91**‡ | .27‡ | .35 | .34† | .11‡ | .18‡ | **.47**† | – | **.54**‡ | **.50**‡ | **.35** | .29 | **.35** |
| BBN-C | **.95**‡ | .21‡ | .22‡ | .36 | .06‡ | .17‡ | **.38** | .26‡ | – | .32 | .24‡ | .31⋆ | .26‡ |
| CMU-HEA-C | **.90**‡ | .17‡ | .19‡ | .23‡ | .09‡ | .18‡ | **.32** | .27‡ | **.34** | – | .31† | .31⋆ | .30‡ |
| JHU-C | **.93**‡ | .19‡ | .30† | .35 | .09‡ | .24‡ | **.50**‡ | .34 | **.47**‡ | **.45**† | – | **.41**‡ | **.36** |
| RWTH-C | **.91**‡ | .16‡ | .35 | .29‡ | .12‡ | .27‡ | **.41**⋆ | **.37** | **.42**⋆ | **.42**⋆ | .23‡ | – | .24† |
| UPV-C | **.94**‡ | .24‡ | **.40** | .36 | .09‡ | .28‡ | **.39** | .32 | **.46**‡ | **.47**‡ | .33 | **.36**† | ? |
| > others | .93 | .26 | .37 | .38 | .11 | .24 | .47 | .40 | .49 | .49 | .38 | .41 | .40 |
| >= others | .97 | .42 | .56 | .55 | .25 | .39 | .67 | .62 | .70 | .70 | .61 | .65 | .62 |

# French-English

| System | constraint? | ≥others |
|---|---|---|
| LIUM ●★ | Y | 0.71 |
| ONLINEB ● | N | 0.71 |
| NRC ●★ | Y | 0.66 |
| CAMBRIDGE ●★ | Y +GW | 0.66 |
| LIMSI ★ | Y +GW | 0.65 |
| UEDIN | Y | 0.65 |
| RALI ●★ | Y +GW | 0.65 |
| JHU | Y | 0.59 |
| RWTH ●★ | Y +GW | 0.55 |
| LIG | Y | 0.53 |
| ONLINEA | N | 0.52 |
| CMU-STATXFER | Y | 0.51 |
| HUICONG | Y | 0.51 |
| DFKI | N | 0.42 |
| GENEVA | Y | 0.27 |
| CU-ZEMAN | Y | 0.21 |

# English-French

| System | constraint? | ≥others |
|---|---|---|
| UEDIN ●★ | Y | 0.70 |
| ONLINEB ● | N | 0.68 |
| RALI ●★ | Y +GW | 0.66 |
| LIMSI ●★ | Y +GW | 0.66 |
| RWTH ●★ | Y +GW | 0.63 |
| CAMBRIDGE ★ | Y +GW | 0.63 |
| LIUM | Y | 0.63 |
| NRC | Y | 0.62 |
| ONLINEA | N | 0.55 |
| JHU | Y | 0.53 |
| DFKI | N | 0.40 |
| GENEVA | Y | 0.35 |
| EU | N | 0.32 |
| CU-ZEMAN | Y | 0.26 |
| KOC | Y | 0.26 |

# German-English

| System | constraint? | ≥others |
|---|---|---|
| ONLINEB ● | N | 0.73 |
| KIT ●★ | Y +GW | 0.72 |
| UMD ●★ | Y | 0.68 |
| UEDIN ★ | Y | 0.66 |
| FBK ★ | Y +GW | 0.66 |
| ONLINEA ● | N | 0.63 |
| RWTH | Y +GW | 0.62 |
| LIU | Y | 0.59 |
| UU-MS | Y | 0.55 |
| JHU | Y | 0.53 |
| LIMSI | Y +GW | 0.52 |
| UPPSALA | Y | 0.51 |
| DFKI | N | 0.50 |
| HUICONG | Y | 0.47 |
| CMU | Y | 0.46 |
| AALTO | Y | 0.42 |
| CU-ZEMAN | Y | 0.36 |
| KOC | Y | 0.23 |

# English-German

| System | constraint? | ≥others |
|---|---|---|
| ONLINEB ● | N | 0.70 |
| DFKI ● | N | 0.62 |
| UEDIN ●⋆ | Y | 0.62 |
| KIT ⋆ | Y | 0.60 |
| ONLINEA | N | 0.59 |
| FBK ⋆ | Y | 0.56 |
| LIU | Y | 0.55 |
| RWTH | Y | 0.51 |
| LIMSI | Y | 0.51 |
| UPPSALA | Y | 0.47 |
| JHU | Y | 0.46 |
| SFU | Y | 0.34 |
| KOC | Y | 0.30 |
| CU-ZEMAN | Y | 0.28 |

# Spanish-English

| System | constraint? | ≥others |
|---|---|---|
| ONLINEB ● | N | 0.70 |
| UEDIN ●⋆ | Y | 0.69 |
| CAMBRIDGE | Y +GW | 0.61 |
| JHU | Y | 0.61 |
| ONLINEA | N | 0.54 |
| UPC ⋆ | Y | 0.51 |
| HUICONG | Y | 0.50 |
| DFKI | N | 0.45 |
| COLUMBIA | Y | 0.45 |
| CU-ZEMAN | Y | 0.27 |

# English-Spanish

| System | constraint? | ≥others |
|---|---|---|
| ONLINEB ● | N | 0.71 |
| ONLINEA ● | N | 0.69 |
| UEDIN ★ | Y | 0.61 |
| DCU | N | 0.61 |
| DFKI ★ | N | 0.55 |
| JHU ★ | Y | 0.55 |
| UPV ★ | Y | 0.55 |
| CAMBRIDGE ★ | Y +GW | 0.54 |
| UHC-UPV ★ | Y | 0.54 |
| SFU | Y | 0.40 |
| CU-ZEMAN | Y | 0.23 |
| KOC | Y | 0.19 |

# Czech-English

| System | constraint? | ≥others |
|---|---|---|
| ONLINEB ● | N | 0.70 |
| UEDIN ⋆ | Y | 0.61 |
| CMU | Y | 0.55 |
| CU-BOJAR | N | 0.55 |
| AALTO | Y | 0.43 |
| ONLINEA | N | 0.37 |
| CU-ZEMAN | Y | 0.22 |

# English-Czech

| System | constraint? | ≥others |
|---|---|---|
| ONLINEB ● | N | 0.70 |
| CU-BOJAR ● | N | 0.66 |
| PC-TRANS ● | N | 0.62 |
| UEDIN ●★ | Y | 0.62 |
| CU-TECTO | Y | 0.60 |
| EUROTRANS | N | 0.54 |
| CU-ZEMAN | Y | 0.50 |
| SFU | Y | 0.45 |
| ONLINEA | N | 0.44 |
| POTSDAM | Y | 0.44 |
| DCU | N | 0.38 |
| KOC | Y | 0.33 |

# Sentence Correction

Ratio of how many edited sentences were judged as correct

| Language pair | Reference | Best system | Best constraint system |
|---|---|---|---|
| French-English | .91 | .58 | .58 |
| English-French | .91 | .54 | .54 |
| German-English | .98 | .80 | .80 |
| English-German | .94 | .80 | .68 |
| Spanish-English | .98 | .71 | .60 |
| English-Spanish | .83 | .58 | .50 |
| Czech-English | 1.00 | .60 | .60 |
| English-Czech | .97 | .58 | .58 |

note: 95% confidence interval is about ±.10

# System Combination Task

- Task: combine output of several systems to produce better translation

- Data provided to participants
  - primary submissions from translation task
  - 25 document subset of submissions along with references as tuning set
  - some systems provided n-best lists

- System combination translations scored alongside individual systems

# Participants

- 8 Institutions
  - Europe: 5
  - North America: 3
  - Asia: 1
- 9 groups
- 41 submitted system translations, also included
  - two popular online translation systems
  - rule-based systems for English–Czech

# Results

- Ranking also includes best individual systems for comparison

- Wins
  - ● indicates a **win** for the system combination meaning that no other system or system combination is statistically significantly better at p-level$\leq$0.1 in pairwise comparison.
  - ⋆ indicates an **individual system** that none of the system combinations beat by a statistically significant margin at p-level$\leq$0.1.

- Note: ONLINEA and ONLINEB were not included among the systems being combined in the system combination shared tasks, except in the Czech-English and English-Czech conditions, where ONLINEB was included.

# French-English

| System | ≥others |
|---|---|
| RWTH-COMBO ● | 0.77 |
| CMU-HYP-COMBO ● | 0.77 |
| DCU-COMBO ● | 0.72 |
| LIUM ⋆ | 0.71 |
| CMU-HEA-COMBO ● | 0.70 |
| UPV-COMBO ● | 0.68 |
| NRC | 0.66 |
| CAMBRIDGE | 0.66 |
| UEDIN ⋆ | 0.65 |
| LIMSI ⋆ | 0.65 |
| JHU-COMBO | 0.65 |
| RALI | 0.65 |
| LIUM-COMBO | 0.64 |
| BBN-COMBO | 0.64 |
| RWTH | 0.55 |

# English-French

| System | ≥others |
|---|---|
| RWTH-COMBO ● | 0.75 |
| CMU-HEA-COMBO ● | 0.74 |
| UEDIN | 0.70 |
| KOC-COMBO ● | 0.68 |
| UPV-COMBO | 0.66 |
| RALI ⋆ | 0.66 |
| LIMSI | 0.66 |
| RWTH | 0.63 |
| CAMBRIDGE | 0.63 |

# German-English

| System | ≥others |
|---|---|
| BBN-COMBO ● | 0.77 |
| RWTH-COMBO ● | 0.75 |
| CMU-HEA-COMBO | 0.73 |
| KIT ⋆ | 0.72 |
| UMD ⋆ | 0.68 |
| JHU-COMBO | 0.67 |
| UEDIN ⋆ | 0.66 |
| FBK | 0.66 |
| CMU-HYP-COMBO | 0.65 |
| UPV-COMBO | 0.64 |
| RWTH | 0.62 |
| KOC-COMBO | 0.59 |

# English-German

| System | ≥others |
|---|---|
| RWTH-COMBO ● | 0.65 |
| DFKI ⋆ | 0.62 |
| UEDIN ⋆ | 0.62 |
| KIT ⋆ | 0.60 |
| CMU-HEA-COMBO ● | 0.59 |
| KOC-COMBO | 0.59 |
| FBK ⋆ | 0.56 |
| UPV-COMBO | 0.55 |

# Czech-English

| System | ≥others |
|---|---|
| CMU-HEA-COMBO ● | 0.71 |
| ONLINEB ⋆ | 0.70 |
| BBN-COMBO ● | 0.70 |
| RWTH-COMBO ● | 0.65 |
| UPV-COMBO ● | 0.63 |
| JHU-COMBO | 0.62 |
| UEDIN | 0.61 |

# English-Czech

| System | ≥others |
|---|---|
| DCU-COMBO ● | 0.75 |
| ONLINEB ★ | 0.70 |
| RWTH-COMBO | 0.70 |
| CMU-HEA-COMBO | 0.69 |
| UPV-COMBO | 0.68 |
| CU-BOJAR | 0.66 |
| KOC-COMBO | 0.66 |
| PC-TRANS | 0.62 |
| UEDIN | 0.62 |

# Spanish-English

| System | ≥others |
|---|---|
| UEDIN ⋆ | 0.69 |
| CMU-HEA-COMBO ● | 0.66 |
| UPV-COMBO ● | 0.66 |
| BBN-COMBO | 0.62 |
| JHU-COMBO | 0.55 |
| UPC | 0.51 |

# English-Spanish

| System | ≥others |
|---|---|
| CMU-HEA-COMBO ● | 0.68 |
| KOC-COMBO | 0.62 |
| UEDIN ⋆ | 0.61 |
| UPV-COMBO | 0.60 |
| RWTH-COMBO | 0.59 |
| DFKI ⋆ | 0.55 |
| JHU | 0.55 |
| UPV | 0.55 |
| CAMBRIDGE ⋆ | 0.54 |
| UPV-NNLM ⋆ | 0.54 |

# Conclusions

- System combinations score better on human judgment
- Most participants were able to use large training corpora
- Mechanical Turk acceptable tool for evaluation