Data Issues of the Multilingual Translation Matrix

Daniel Zeman
Charles University in Prague, Institute of Formal and Applied Linguistics
Univerzita Karlova v Praze, Ústav formální a aplikované lingvistiky
Malostranské náměstí 25, CZ-11800 Praha
zeman@ufal.mff.cuni.cz

PARALLEL DATA

Corpus	SentPairs	Tokens lng1	Tokens lng2
cs-en	782,756	17,997,673	20,964,639
de-en	2,079,049	55,143,719	57,741,141
es-en	2,123,036	61,784,972	59,217,471
fr-en	2,144,820	69,568,241	59,939,548
de-cs	652,193	17,422,620	15,383,601
es-cs	692,118	20,189,811	16,324,910
fr-cs	686,300	22,220,780	16,190,365
un.es-en	11,196,913	368,154,702	328,824,317
un.fr-en	12,886,831	449,279,647	372,627,886

Table 1: Number of sentence pairs and tokens for every language pair in the parallel training corpus (EuroParl + News Commentary v7; last two lines are the United Nations corpora). The xx-cs corpora were obtained as intersections of xx-en and cs-en.

MONOLINGUAL DATA

Corpus	Segments	Tokens
newsc+euro.cs	819,434	18,491,692
newsc+euro.de	2,360,811	58,683,607
newsc+euro.en	2,430,718	65,934,441
newsc+euro.es	2,307,429	66,072,443
newsc+euro.fr	2,361,764	74,083,166
news.all.cs	14,552,899	244,728,011
news.all.de	24,446,319	462,924,303
news.all.en	42,161,804	1,039,806,242
news.all.es	8,627,438	249,022,213
news.all.fr	16,708,622	438,489,352
gigaword.en	70,592,779	2,546,581,646
gigaword.es	31,304,148	1,064,660,498
gigaword.fr	21,674,453	963,571,174

Table 2: Number of segments (paragraphs or sentences) and tokens for every monolingual corpus. Newsc = News Commentary; news.all = crawled news from all the years.

BLEU

Direction	Baseline	news.all	gigaword
en-cs	0.1196	0.1434	
en-de	0.1426	0.1629	
en-es	0.2778	0.3136	0.3136
en-fr	0.2599	0.2897	0.2874
cs-en	0.1796	0.2031	0.2013
de-en	0.1877	0.2136	0.2144
es-en	0.2219	0.2428	0.2390
fr-en	0.2459	0.2764	0.2756
cs-de	0.1365	0.1550	
cs-es	0.1952	0.2211	0.2184
cs-fr	0.1953	0.2167	0.2147
de-cs	0.1212	0.1400	
es-cs	0.1281	0.1489	
fr-cs	0.1253	0.1442	

Table 3: BLEU scores. Baseline: LM on parallel corpus only. Other columns add news.all or gigaword respectively.

UN CORPUS

Direction	Parallel	Mono	BLEU
en-es	news-euro-un	news.all	0.3194
en-es	news-euro	news.all	0.3136
en-es	un	un	0.2694
en-fr	news-euro	news.all	0.2897
en-fr	un	un	0.2541
es-en	un	un	0.2688
es-en	news-euro	news.all	0.2428
fr-en	news-euro	news.all	0.2764
fr-en	un	un	0.2392

Table 4: BLEU scores with and without the UN corpus.

HUMAN

Direction	BLEU	Hwin
en-fr	0.2897	0.549350
en-de	0.1629	0.545068
fr-en	0.2764	0.534239
en-cs	0.1434	0.533774
en-es	0.3136	0.525881
de-en	0.2144	0.499220
cs-en	0.2031	0.463432
es-en	0.2428	0.259303

Table 5: Human comparison with other systems (percentage of times when this system produced better translation than another system, ignoring ties).

THE SYSTEM

- Tokenization, in-house quotation marks normalization, supervised truecasing (based on lemmas ← TreeTagger or Morče)
- Word alignment: Giza++ operating on lemmas
- Hexagram language model
- Moses decoder, no lexical reordering model, no factored model
- All BLEU scores computed by the system on tokenized test data, truecased (except for the sentence-initial letter)

Supervised truecasing
Training data with lemma factor:
A token is lowercased unless its lemma is uppercased.

Changes sentence-inital tokens, words in English headings, highlighting uppercase etc. *Errors on common-proper ambiguities?*

mušaraf	+ův	po+ sled +ní	výstup	?				
musharraf	s	last	act	?				
udusit	ne+	závislé	soud + nictví	a	svobod +ná	média	.	
to stifle	in+	dependent	judiciary	and	free	media	.	
nic ne+	může	být	dále	od	pravd +y	.		
nothing	could	be	further	from	the truth	.		
mnoho	pákistán	+ců	rozčar +ova	+ných	pákistán	+skou	po+ liticko +u	garnit +u + rou
many	pakistan	+is	dis+ illusion	+ed	with pakistan	s	political	class

Unsupervised morphemic segmentation

en-cs: Swapped words

en-cs: Wrong case

cs-es: Bad word order

Direction	Translation
en	**The story of Libya's liberation, or rebellion, already has its defeated.**
en-cs	Příběh Libye osvobození, nebo odboj, už svou porážku.
en-de	Die Geschichte von Libyens Befreiung oder Rebellion, hat bereits seine Niederlage.
en-es	La historia de la liberación de Libia, o la rebelión, ya tiene su derrotado.
en-fr	L'histoire de la libération de la Libye, ou la rébellion, a déjà son vaincu.
cs	**Příběh libyjského osvobození, anebo libyjské rebélie, již má své poražené.**
cs-de	Die Geschichte des libyschen Befreiung, oder der libyschen rebélie, hat bereits seine Verlierer.
cs-en	The story of the libyan liberation or libyan rebélie, already has its losers.
cs-es	La historia de la liberación de Libia, Libia o rebélie, ya tiene perdedores.
cs-fr	L'histoire de la libération de la Libye, ou rebélie, possède déjà des perdants.
de	**Die Geschichte des libyschen Befreiungskampfes oder libyschen Rebellion kennt schon ihre Verlierer.**
de-cs	Historie libyjského Befreiungskampfes nebo libyjského povstání už zná své poražené.
de-en	The history of the libyan liberation or libyan rebellion knows about their losers.
es	**La historia de la liberación libia, o de la rebelión libia, ya tiene sus perdedores.**
es-cs	Historie osvobození Libye, nebo libyjské povstání, protože má své poražené.
es-en	The history of the liberation of Libya, or of the rebellion in Libya, already has sus perdedores.
es-en.un	The history of liberation libyan, or of the rebellion libyan, already has its losers.
fr	**La libération ou rebellion libyenne a déjà ses vaincus.**
fr-cs	Propuštění nebo rebellion libyjské již své poražené.
fr-en	The release or libyan rebellion already has its losers.

en-cs & en-de: “defeated” mistranslated as “defeat”

cs: “rebélie” is very rare, synonyms exist

en-cs & fr-cs: missing verb “has” (often dropped when translating perfect tense)

fr-en & fr-cs: “libération” wrongly translated as “release” instead of “liberation”

es-en: The UN corpus has larger vocabulary (“perdedores”) but it also prefers strangely ordered N-Adj phrases.

Example 1: Reference translations and system hypotheses for one sentence. Original language is Czech.

French Source
le parlement n ' a pas ratifié l ' amendement pour la libération de Tymosenko
le parlement ukrainien a refusé , dans la cadre d ' un amendement au droit pénal , le projet d ' annulation du paragraphe relatif à l ' inculpation de Julia Tymosenko , la chef de l ' opposition .
les députés ont refusé en deuxième lecture le projet de modification visant la réduction des peines pour délits économiques , qui aurait pu ouvrir les portes de la liberté pour l ' ex - Première Ministre actuellement emprisonnée .
Tymosenko a été condamnée en octobre à 7 ans de prison pour avoir conclu un accord à priori désavantageux avec la Russie pour l ' achat de gaz naturel .
le jugement n ' est pas définitif et le tribunal doit statuer sur l ' appel de la condamnée en décembre .
Tymosenko qualifie le jugement de vengeance politique du régime et a provoqué un processus de soupçons de partialité du tribunal également à l ' ouest .
le projet d ' annuler le paragraphe 365 du droit pénal , sur la base duquel l ' ex - Premère Ministre a été condamnée , a été signé par 147 députés .
il aurait fallu 226 voix pour l ' approuver .

English Reference
Parliament Does not Support Amendment Freeing Tymoshenko
today , the Ukraine parliament dismissed , within the Code of Criminal Procedure amendment , the motion to revoke an article based on which the opposition leader , Yulia Tymoshenko , was sentenced .
the amendment that would lead to freeing the imprisoned former Prime Minister was revoked during second reading of the proposal for mitigation of sentences for economic offences .
in October , Tymoshenko was sentenced to seven years in prison for entering into what was reported to be a disadvantageous gas deal with Russia .
the verdict is not yet final ; the court will hear Tymoshenko ' s appeal in December .
Tymoshenko claims the verdict is a political revenge of the regime ; in the West , the trial has also evoked suspicion of being biased .
the proposal to remove Article 365 from the Code of Criminal Procedure , upon which the former Prime Minister was sentenced , was supported by 147 members of parliament .
its ratification would require 226 votes .

Translation Trained on News Commentary + Europarl; LM from Crawled News.ALL
Parliament has not ratified the amendment for the release of Tymosenko
the ukrainian parliament has refused , in the context of a change in the criminal law , the cancellation of the paragraph relating to the prosecution of Julia Tymosenko , the leader of the opposition .
members have refused at second reading the draft amendment to reduce the penalties for economic crimes , which could open the doors of freedom for the former First Minister currently in prison .
Tymosenko was sentenced in October to 7 years in prison for having reached an agreement in principle with Russia disadvantageous for the purchase of natural gas .
the verdict is not final , and the court must rule on the appeal of the sentence in December .
Tymosenko describes the judgment of political retribution by the regime and led to a process of suspected bias of the court , also to the west .
the project to cancel the (365 of criminal law , on the basis of which the former Premère Minister was sentenced , has been signed by 147 members .
it had 226 votes for approval .

Language Model Trained on English Gigaword: slightly better despite BLEU?
the parliament has not ratified the amendment for the release of Tymosenko
the ukrainian parliament has refused , in the framework of an amendment to the criminal law , the cancellation of a section in charge of Julia Tymosenko , the leader of the opposition .
the members have refused the second reading of the draft amendment to reduce the penalties for economic crimes , which could open the doors of freedom for the former First Minister currently in prison .
Tymosenko was sentenced in October to seven years in prison for having reached an agreement at a disadvantage with Russia for the purchase of natural gas .
the verdict is not final and the court must rule on the appeal of the convicted in December .
Tymosenko called the judgment of political retribution by the regime and has provoked a process of suspected partiality of the court also in the west .
the project to cancel the (365 of the criminal law , on the basis of which the former Premère minister was sentenced , was signed by 147 members .
it would have required 226 votes for approval .

United Nations Corpus (both Translation and Language Model): not so bad at all…
the Parliament has not ratified the amendment for the release of Tymosenko
the Ukrainian Parliament refused , in the framework of an amendment to the criminal law , the proposed cancellation of the paragraph concerning the indictment of Julia Tymosenko , the leader of the opposition .
the deputies have refused in second reading the draft amendment the reduction of penalties for economic crimes , which could open the doors of freedom for the Former Prime Minister currently imprisoned .
Tymosenko was ordered in October to 7 years in prison for having concluded an agreement a priori disadvantageous with Russia for the purchase of natural gas .
the judgement is not final and the court shall decide on the appeal of the sentenced in December .
Tymosenko qualifies the judgement of political revenge of the regime and has led to a process of suspicion of bias of the court also in the west .
the draft annul paragraph 365 of the Penal Law , on the basis of which the former Premère Minister was sentenced , was signed by 147 deputies .
it should have been 226 votes to approve .

The research has been supported by the grant P406/11/1499.