Click to edit Master subtitle style

Gennadi Lembersky
Noam Ordan
Shuly Wintner
MTML, 2011

Background

Background: Original vs. Translated Texts

Soure Text $\xrightarrow{\text{LM}}$ **Target Text**

Soure Text $\xleftrightarrow{\text{TM}}$ **Target Text**

Wanderer's Night Song Up there all summits are still. In all the tree-tops you will feel but the dew. The birds in the forest stopped talking. Soon, done with walking, you shall rest, too. (~50 translations into Hebrew)	Wandrers Nachtlied Über allen Gipfeln ist Ruh, in allen Wipfeln spürest du kaum einen Hauch; die Vögelein schweigen im Walde, warte nur, balde ruhest du auch! (26 tokens)

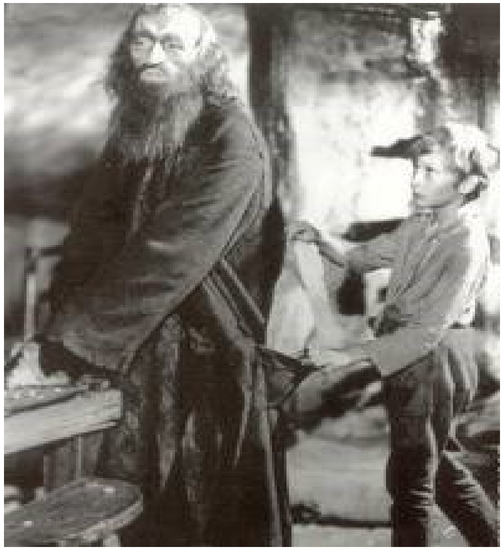
Background: Is sex/translation dirty?

LM

Soure Text

TM

Target Text



Background: Original vs. Translated Texts

Given this simplified model:

Two points are made with regard to the "intermediate component" (TM and LM):

1. TM is blind to direction (but see Kurokawa et al., 2009)
2. LMs are based on originally written texts.

Background: Original vs. Translated Texts

LMs are based on originally written texts for two possible reasons:

1. They are more readily available;
2. Perhaps the question of whether they are translated or not is considered irrelevant for LM.

Background: Original vs. Translated Texts

Translated texts are ontologically different from non-translated texts ; they generally exhibit

1. ***Simplification*** *of the message, the grammar or both* (Al-Shabab, 1996, Laviosa, 1998) ;
2. ***Explicitation*****,** the tendency to spell out implicit utterances that occur in the source text (Blum-Kulka, 1986).

Background: Original vs. Translated Texts

- Translated texts can be distinguished from non-translated texts with high accuracy (87% and more)
  - For Italian (Baroni & Bernardini, 2006)
  - For Spanish (Iliseiet al., 2010);
  - For English (Koppel & Ordan, forthcoming)

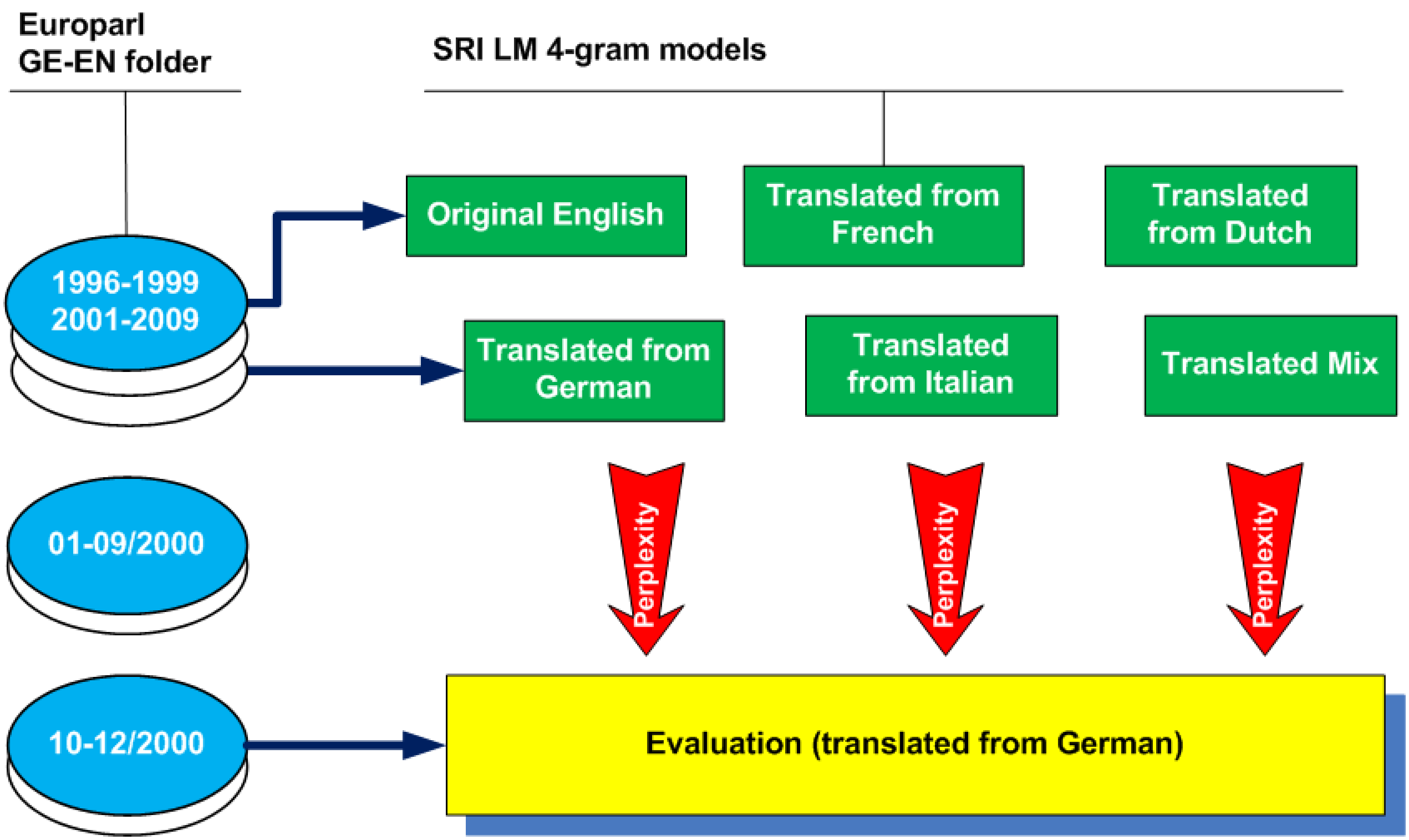
Hypotheses

Our Hypotheses

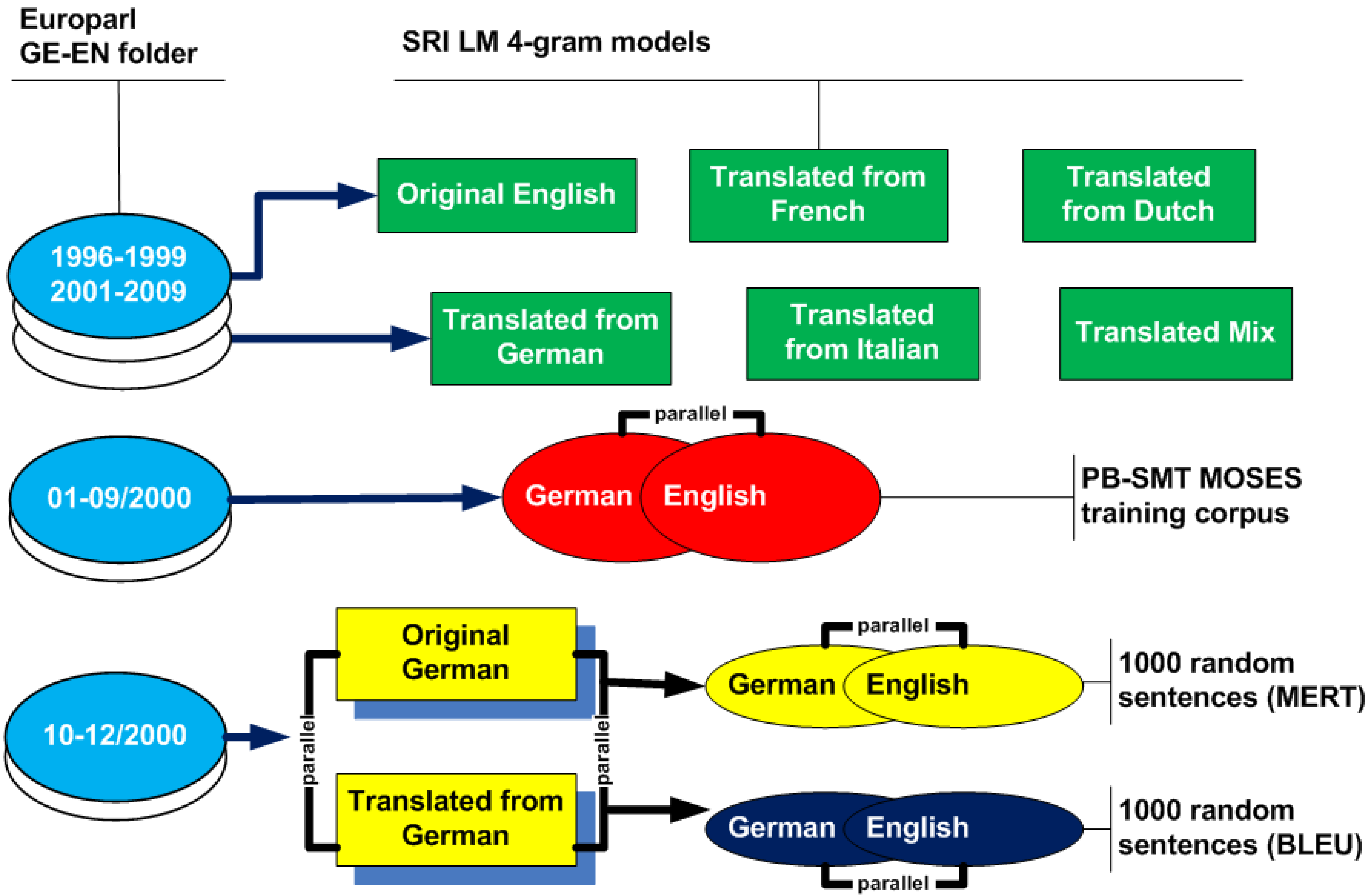
We investigate the following three hypotheses:

1. Translated texts differ from original texts
2. Texts translated from one language differ from texts translated from other languages
3. LMs compiled from translated texts are better for MT than LMs compiled from original texts

Testing Hypothesis 1+2



Testing Hypothesis 3



Identifying the Source Language

- For the most part, we rely on the LANGUAGE attribute of the SPEAKER tag
  - <SPEAKER LANGUAGE="DE" ID="..."/>
  - BUT: it is rarely used with British MEPs
- To identify original English speakers we use ID attribute, which we match against the list of British members of the European parliament

Europarl Experiments

Resources

- 4 European language pairs taken from Europarl
  - German – English
  - Dutch – English
  - French – English
  - Italian – English

Language Models Stats

German - English			
Len	**Tokens**	**Sent's**	**Orig. Lang.**
28.12	2,325,261	82,700	Mix
25.52	2,324,745	91,100	O-EN
26.43	2,322,973	87,900	T-DE
24.72	2,323,646	94,000	T-NL
29.98	2,325,183	77,550	T-FR
35.68	2,325,996	65,199	T-IT

Dutch - English			
Len	**Tokens**	**Sent's**	**Orig. Lang.**
27.72	2,508,265	90,500	Mix
25.52	2,475,652	97,000	O-EN
26.57	2,503,354	94,200	T-DE
24.66	2,513,769	101,950	T-NL
29.13	2,523,055	86,600	T-FR
34.24	2,518,196	73,541	T-IT

Language Models Stats

French - English			
Len	**Tokens**	**Sent's**	**Orig. Lang.**
28.07	2,546,274	90,700	Mix
25.64	2,545,891	99,300	O-EN
26.83	2,546,124	94,900	T-DE
24.63	2,545,645	103,350	T-NL
29.69	2,546,085	85,750	T-FR
35.37	2,546,984	72,008	T-IT

Italian - English			
Len	**Tokens**	**Sent's**	**Orig. Lang.**
29.12	2,534,793	87,040	Mix
27.11	2,534,892	93,520	O-EN
27.99	2,534,867	90,550	T-DE
26.18	2,535,053	96,850	T-NL
30.57	2,534,930	82,930	T-FR
36.60	2,535,225	69,270	T-IT

SMT Training Data

Len	Tokens	Sent's	Side	Lang's
26.26	2,439,370	92,901	DE	DE-EN
28.01	2,602,376	92,901	EN	
27.44	2,327,601	84,811	NL	NL-EN
27.16	2,303,846	84,811	EN	
28.02	2,610,551	93,162	FR	FR-EN
30.80	2,869,328	93,162	EN	
29.62	2,531,925	85,485	IT	IT-EN
29.45	2,517,128	85,485	EN	

Reference Sets

Len	Tokens	Sent's	Side	Lang's
24.25	161,889	6,675	DE	DE-EN
26.81	178,984	6,675	EN	
24.88	114,272	4,593	NL	NL-EN
22.88	105,083	4,593	EN	
30.63	260,198	8,494	FR	FR-EN
31.97	271,536	8,494	EN	
36.25	82,261	2,269	IT	IT-EN
34.49	78,258	2,269	EN	

Hypotheses 1+2 Results

German - English		
PP	**Unigrams**	**Orig. Lang.**
83.45	32,238	Mix
96.50	31,204	O-EN
77.77	27,940	T-DE
89.17	28,074	T-NL
92.71	29,405	T-FR
95.14	28,586	T-IT

Dutch - English		
PP	**Unigrams**	**Orig. Lang.**
87.37	33,050	Mix
100.75	32,064	O-EN
90.35	28,766	T-DE
78.25	29,178	T-NL
96.38	30,502	T-FR
99.26	29,386	T-IT

Hypotheses 1+2 Results

French - English		
PP	**Unigrams**	**Orig. Lang.**
87.13	33,444	Mix
105.93	32,576	O-EN
96.83	28,935	T-DE
100.18	29,221	T-NL
82.23	30,609	T-FR
91.15	29,633	T-IT

Italian - English		
PP	**Unigrams**	**Orig. Lang.**
90.71	33,353	Mix
107.45	32,546	O-EN
100.46	28,835	T-DE
105.07	29,130	T-NL
92.18	30,460	T-FR
80.57	29,466	T-IT

Hypothesis 1+2 Results

- Corpora statistics and LM perplexity results support the hypotheses:
  - translated and original texts are different
  - texts translated from one language are different from texts translated from another language
- For every source language, L:
  - LM trained on texts translated from L has the lowest (the best) perplexity
  - The MIX LMs are second-best and the LMs trained on texts translated from related languages (German<->Dutch; French<->Italian) are next
  - The LMs trained on original English texts are the worst

Hypotheses 3 (MT) Results

German - English	
BLEU	**Orig. Lang**
21.95	Mix
21.35	O-EN
22.42	T-DE
21.59	T-NL
21.47	T-FR
21.79	T-IT

Dutch - English	
BLEU	**Orig. Lang**
25.17	Mix
24.46	O-EN
25.12	T-DE
25.73	T-NL
24.79	T-FR
24.93	T-IT

French - English	
BLEU	**Orig. Lang**
25.43	Mix
24.85	O-EN
25.03	T-DE
25.17	T-NL
25.91	T-FR
25.44	T-IT

Italian - English	
BLEU	**Orig. Lang**
26.79	Mix
25.69	O-EN
25.86	T-DE
25.77	T-NL
26.56	T-FR
27.28	T-IT

Hypotheses 3 (MT) Results / 2

- The results support the hypothesis:
  - For every source language L, the MT system that uses LM trained on text translated from L has the best translations.
  - Systems that use O-EN LMs got the lowest BLEU scores.
- Statistical significance (bootstrap resampling):
  - The best-performing system is statistically better than all other systems ($p < 0.05$)
  - The best-performing system is statistically better than O-EN system ($p < 0.01$)
  - The MIX systems are statistically better than O-

Hebrew-English Experiments

Hebrew-English MT System

- MOSES PB-SMT
- Factored Translation Model (surface | lemma) trained on ~ 65,000 parallel sentences
- Fully segmented source (Hebrew)
  - Morphological analyzer (from "MILA" knowledge center) and Roy Bar-Haim's disambiguator
- Lemma-based alignment + "trgtosrc alignment"
- Performance:
  - ~ 23 BLEU on 1000 sentences with 1 ref. translations
  - ~ 32 BLEU on 300 sentences with 4 ref.

Language Model Resources

- Two English Corpora for the language models
  - **Original English corpus (O-EN)** – “International Herald Tribune” articles collected over a period of 7 months (January to July 2009)
  - **Translated from Hebrew (T-HE)** – Israeli newspaper “HaAretz” published in Hebrew collected over the same period of time
- Each corpus comprises 4 topics: news, business, opinion and arts
  - Both corpora have approximately the same number of tokens in each topic

Language Models Resources

Hebrew - English			
Len	**Tokens**	**Sent's**	**Orig. Lang.**
26.3	3,561,559	135,228	O-EN
24.2	3,561,556	147,227	T-HE

Parallel Resources

- SMT Training Model
  - Hebrew-English parallel corpus (Tsvetkov and Wintner, 2010)
    - Genres: news, literature and subtitles
    - Original Hebrew (54%)
    - Original English (46%) – mostly subtitles
- Reference Set
  - Translated from Hebrew to English
  - Literature (88.6%) and news (11.4%)

Parallel Resources

Len	Tokens	Sent's	Side	Lang's
SMT Training Data				
7.6	726,512	95,912	HE	HE-EN
8.9	856,830	95,912	EN	
Reference Set				
13.5	102,085	7,546	HE	HE-EN
16.7	126,183	7,546	EN	

Hypothesis 1 Results

Hebrew - English		
PP	**Unigrams**	**Orig. Lang.**
282.75	74,305	O-EN
226.02	61,729	T-HE

- **Problem**: What if the different perplexity results are due to the contents bias between T-HE corpus and the reference sets
  - We conducted more experiments in which we gradually abstract away from the specific contents

Abstraction Experiments

- 4 abstraction levels:
  - 1 – we remove all punctuation
  - 2 – we replace named entities with a “NE” token
    - We use Stanford Named Entity Recognizer
    - We train 5-gram LMs
  - 3 – we replace all nouns with a their POS tag
    - We use Stanford POS Tagger
    - We train 5-gram LMs
  - 4 – we replace all tokens with their POS tags
    - We train 8-gram LM

Abstraction Experiments

PP diff.	T-HE	O-EN	Abstraction
	PP	PP	
19.2%	358.11	442.95	No Punctuation
17.3%	289.71	350.3	NE Abstraction
12.4%	81.72	93.31	Noun Abstraction
6.2%	10.76	11.47	POS Abstraction

- T-HE fits the reference consistently better than O-EN

Hypothesis 3 (MT) Results

Hebrew - English	
BLEU	**Orig. Lang**
11.98	O-EN
12.57	T-HE

- T-HE system produces slightly better results
- The gain is statistically significant ($p = 0.012 < 0.05$)

Discussion

Discussion

The results consistently support our hypotheses:

1. Translated texts differ from original texts
2. Texts translated from one language differ from texts translated from other languages
3. LMs compiled from translated texts are better for MT than LMs compiled from original texts

Discussion

Practical Outcome:

- Use LMs trained on texts translated from a source language
- If not available, use the mixture of translated texts
  - The texts translated from languages closely-related to the source language are for most part better than other translated texts

Discussion

Why did it work? Two hypotheses:

1. Since translations simplify the originals, error potential gets smaller and LMs better predict translated language;
2. Recurrent multiword expressions in the SL converge to a set of high-quality translations in the TL.

Discussion

When machine translation meets translation studies,

1. MT Results improve;
2. Pending hypotheses in translation studies are tested experimentally in a more rigorous way.

We call for further cooperation.

Thank You!