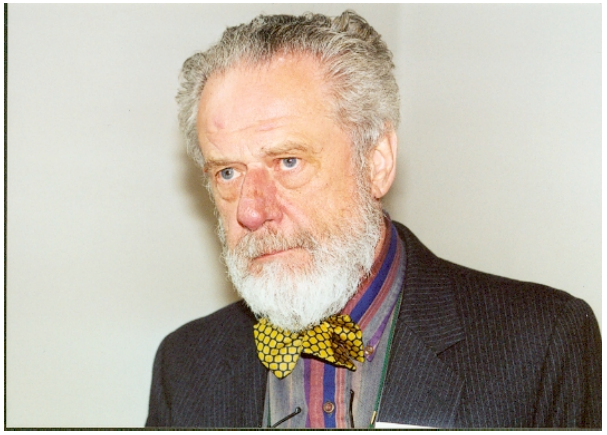
MACHINE TRANSLATION PAPER 1

Daniel Montalvo, Chrysanthia Cheung-Lau, Jonny Wang
CS159 Spring 2011

Paper

- Intersecting multilingual data for faster and better statistical translations. Association for Computational Linguistics, June 2009.

Yu Chen



Martin Kay

Andreas Eisele

Introduction

- Statistical Machine Translation (SMT):

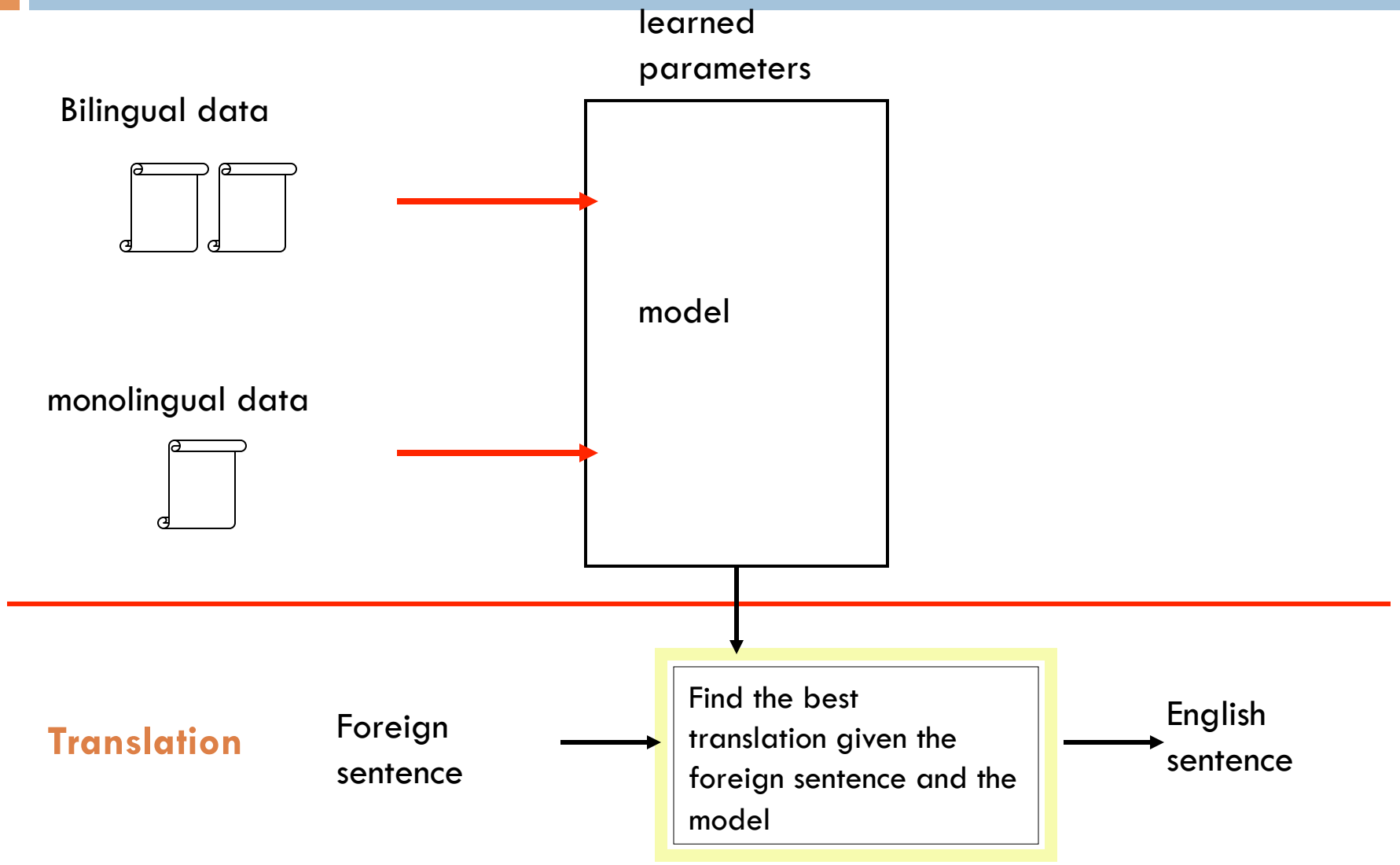
model

$$p(e \mid f) \propto p(f \mid e)\, p(e)$$

translation model

language model

Statistical MT Overview



Problems

- Noise from word/phrase alignment
- Efficiency: time and space

Problems

Sie lieben ihre Kinder nicht

They love their children not

They don't love their children

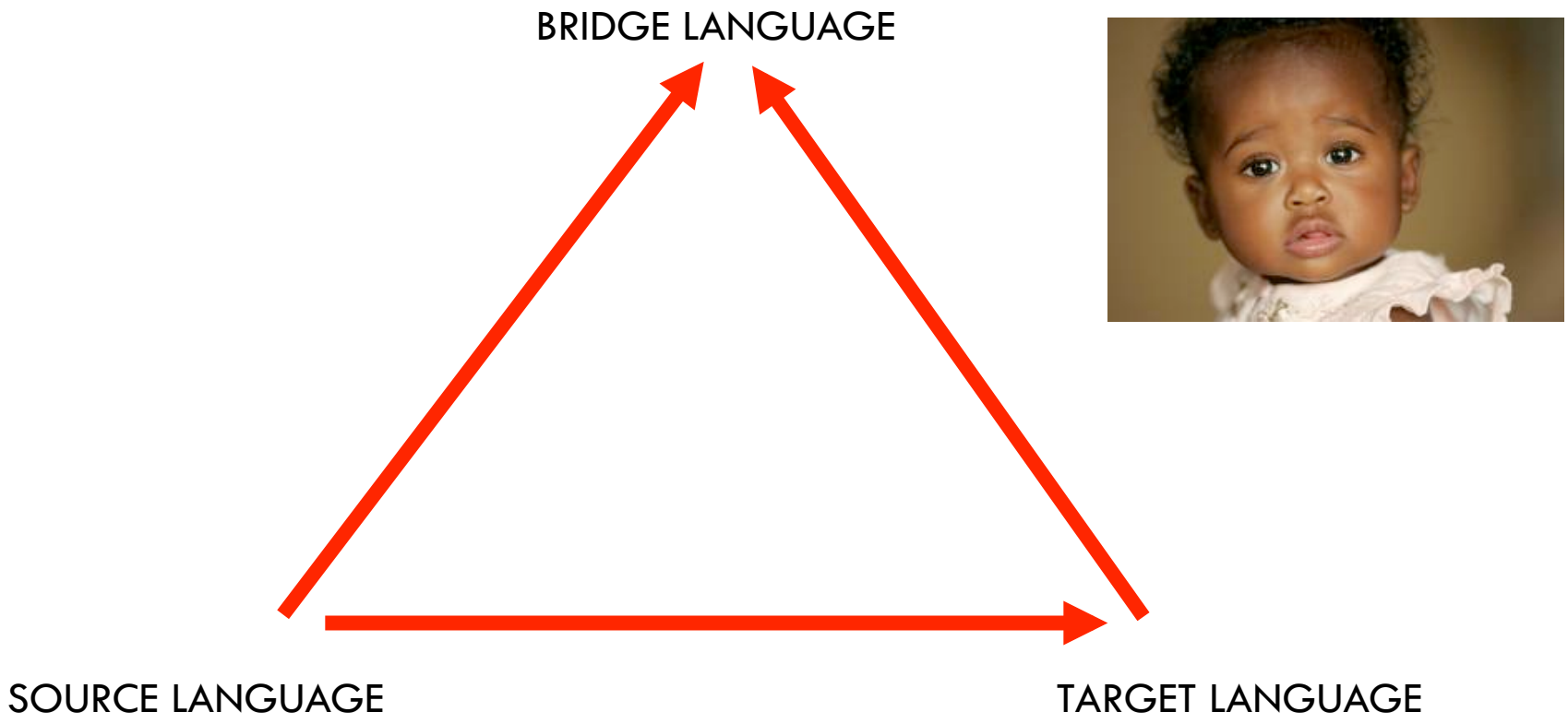
ihre Kinder nicht => a) their children are not b) their children

Language model =/= They love their children are not

Solution: remove b) from translation model

- http://www.translationparty.com/#9144687
- http://www.translationparty.com/#9144706
- http://www.translationparty.com/#9144869
- http://www.translationparty.com/#9144872
- http://www.translationparty.com/#9144881

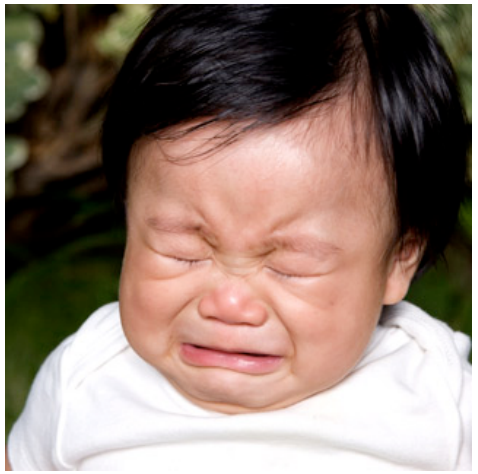
Solution: Triangulation



Triangulation

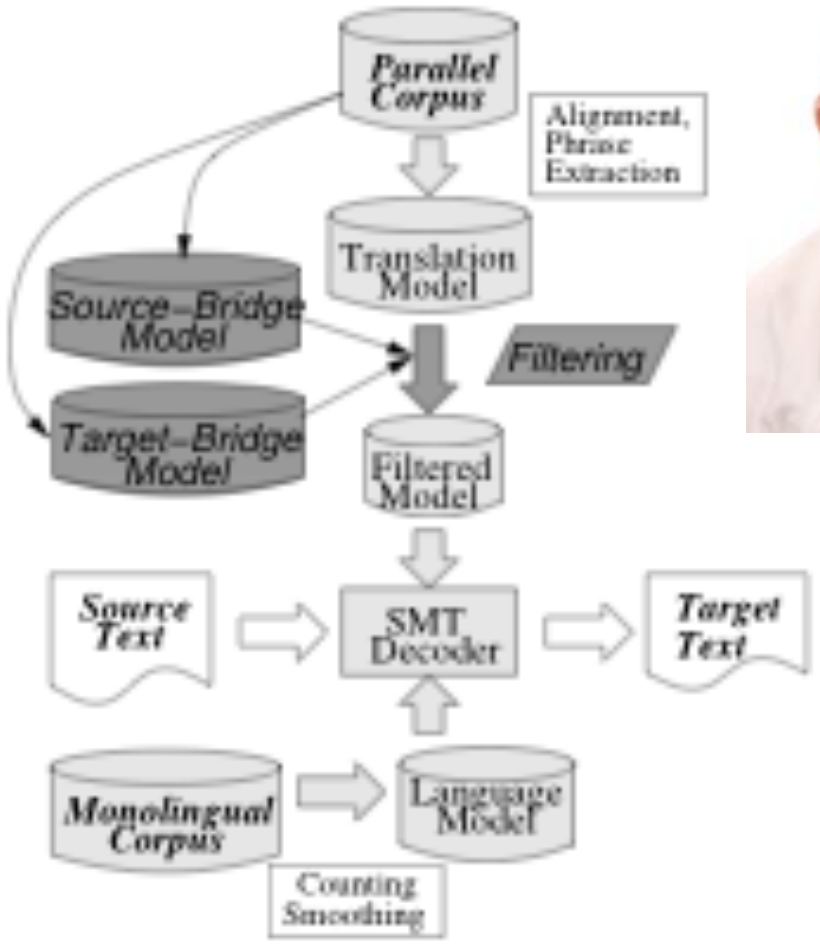
- Look at source-target phrase pair (S,T)

- S => X, T => X, S => T (keep)
- S => X, T => Y, S =/> T (drop)
- S => ?, T => ?, ??? (keep)



Experimental Design

- Database: Europarl (1.3 million)
- Divided by maximal sentence length
- Spanish => English; bridges: French, Portuguese, Danish, German, Finnish
- Models from Moses toolkit
- Train, develop, test



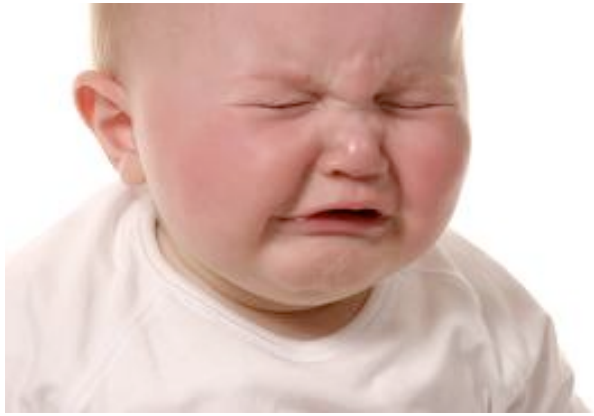
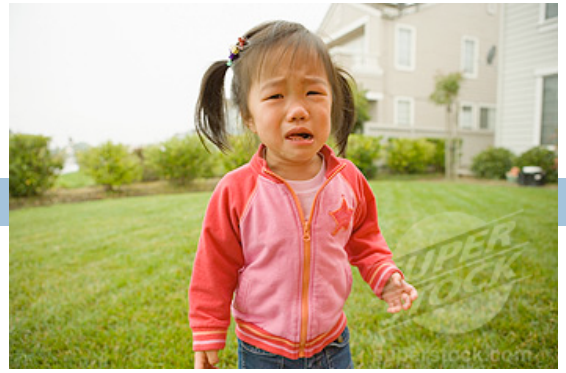


Figure 1: Triangulated filtering in SMT systems

Results



Bridge	EP-20	EP-40	EP-50
—	26.62	31.43	31.68
pt	28.40	**32.90**	**33.93**
fr	28.28	32.69	33.47
da	**28.48**	32.47	33.88
de	28.05	32.65	33.13
fi	**28.02**	**31.91**	**33.04**

Table 3: BLEU scores of translations using filtered phrase tables

Model+Bridge	Time		Table Size		
	T_l (s)	T_t (s)	N	S_{PT} (byte)	S_{RT} (byte)
EP-20+ —	55	3529	7,599,271	953M	717M
EP-20+ pt	**53**	**2826**	**1,712,508** (22.54%)	**198M**	**149M**
EP-20+ fr	48	2702	1,536,056 (20.21%)	172M	131M
EP-20+ da	52	2786	1,659,067 (21.83%)	186M	141M
EP-20+ de	**43**	2732	**1,260,524** (16.59%)	**132M**	**101M**
EP-20+ fi	47	**2670**	1,331,323 (17.52%)	147M	111M
EP-40+ —	65	3673	19,199,807	2.5G	1.9G
EP-40+ pt	**50**	**3091**	8,378,517 (43.64%)	1.1G	1.8G
EP-40+ fr	46	3129	**8,599,708** (44.79%)	**1.1G**	**741M**
EP-40+ da	42	3050	6,716,304 (34.98%)	842M	568M
EP-40+ de	46	3069	6,113,769 (31.84%)	725M	492M
EP-40+ fi	**40**	**2889**	**4,473,483** (23.30%)	**533M**	**353M**
EP-50+ —	140	4130	54,382,715	7.1G	5.4G
EP-50+ pt	78	3410	13,225,654 (24.32%)	1.6G	1.3G
EP-50+ fr	**97**	**3616**	**24,057,849** (44.24%)	**3.0G**	**2.3G**
EP-50+ da	81	3418	12,547,839 (23.07%)	1.5G	1.2G
EP-50+ de	95	3488	15,938,151 (29.31%)	1.9G	1.5G
EP-50+ fi	**71**	**3191**	**7,691,904** (17.75%)	**895M**	**677M**

Table 2: System efficiency: time consumption and phrase-table size

Phrase Length

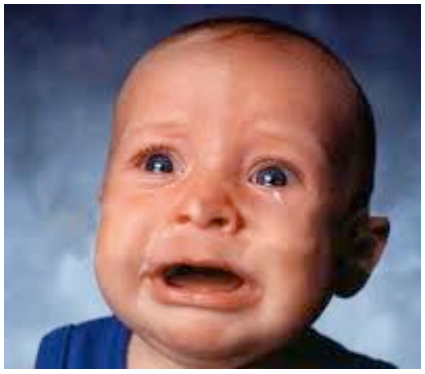
- "Long phrases suck" (Koehn, 2003)

fabricantes	pt	fr	da	de	fi	
a manufacturer	✓	✓	✓		✓	4
battalions	✓	✓	✓			3
car manufacturers have						0
car manufacturers	✓	✓	✓	✓	✓	5
makers	✓	✓			✓	3
manufacturer	✓	✓	✓	✓	✓	5
manufacturers	✓	✓	✓	✓	✓	5
producers are		✓	✓	✓		3
producers need						0
producers	✓	✓	✓	✓	✓	5

Table 6: Phrase-table entries before and after filtering a model with different bridges

Bridge	EP-20	EP-40	EP-50
—	3.776	4.242	4.335
pt	3.195	3.943	3.740
fr	3.003	3.809	3.947
da	3.005	3.74	3.453
de	2.535	3.501	3.617
fi	2.893	3.521	3.262

Table 5: Average phrase length



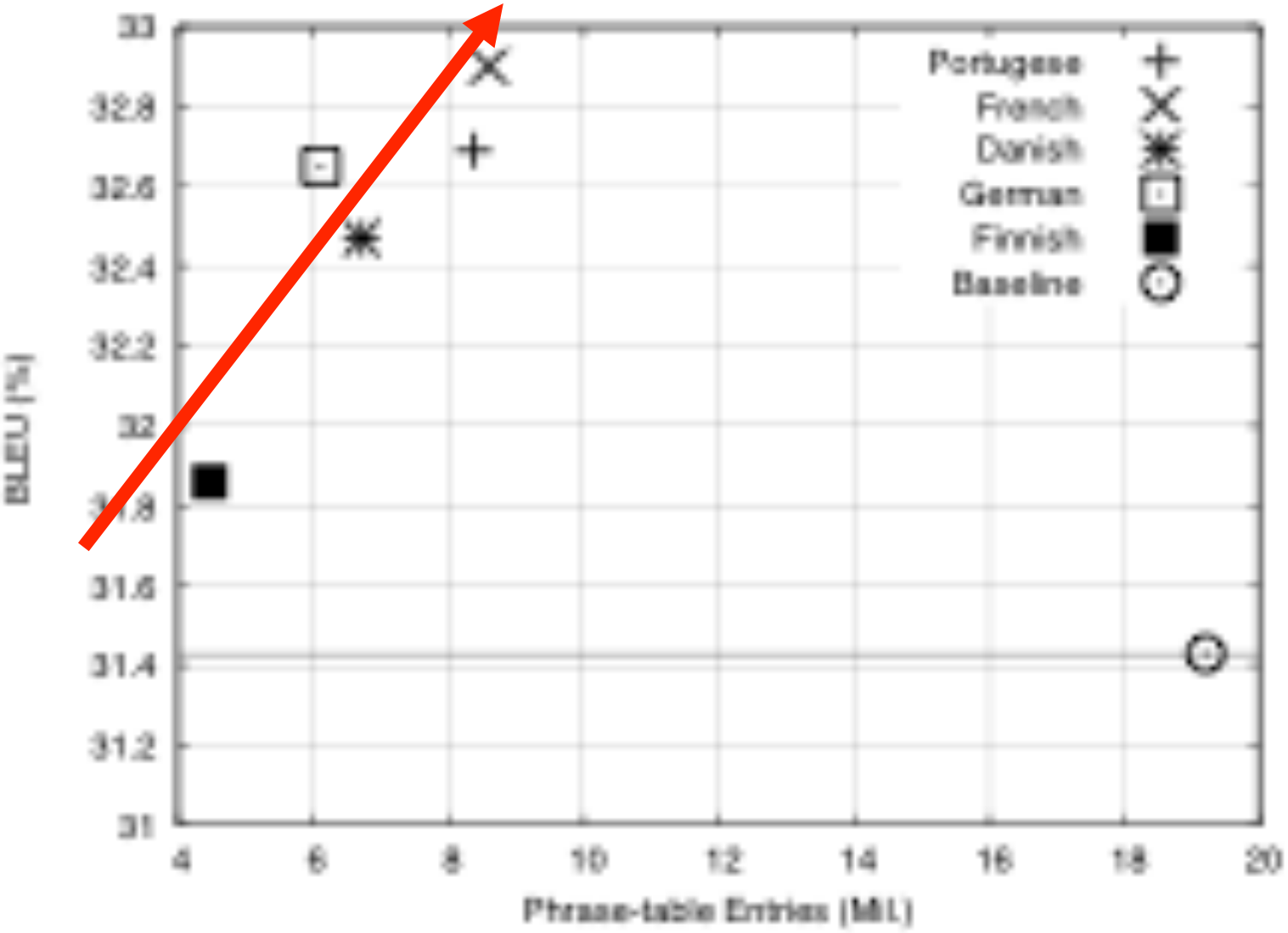


Figure 2: Clustering of bridge languages

Further Research

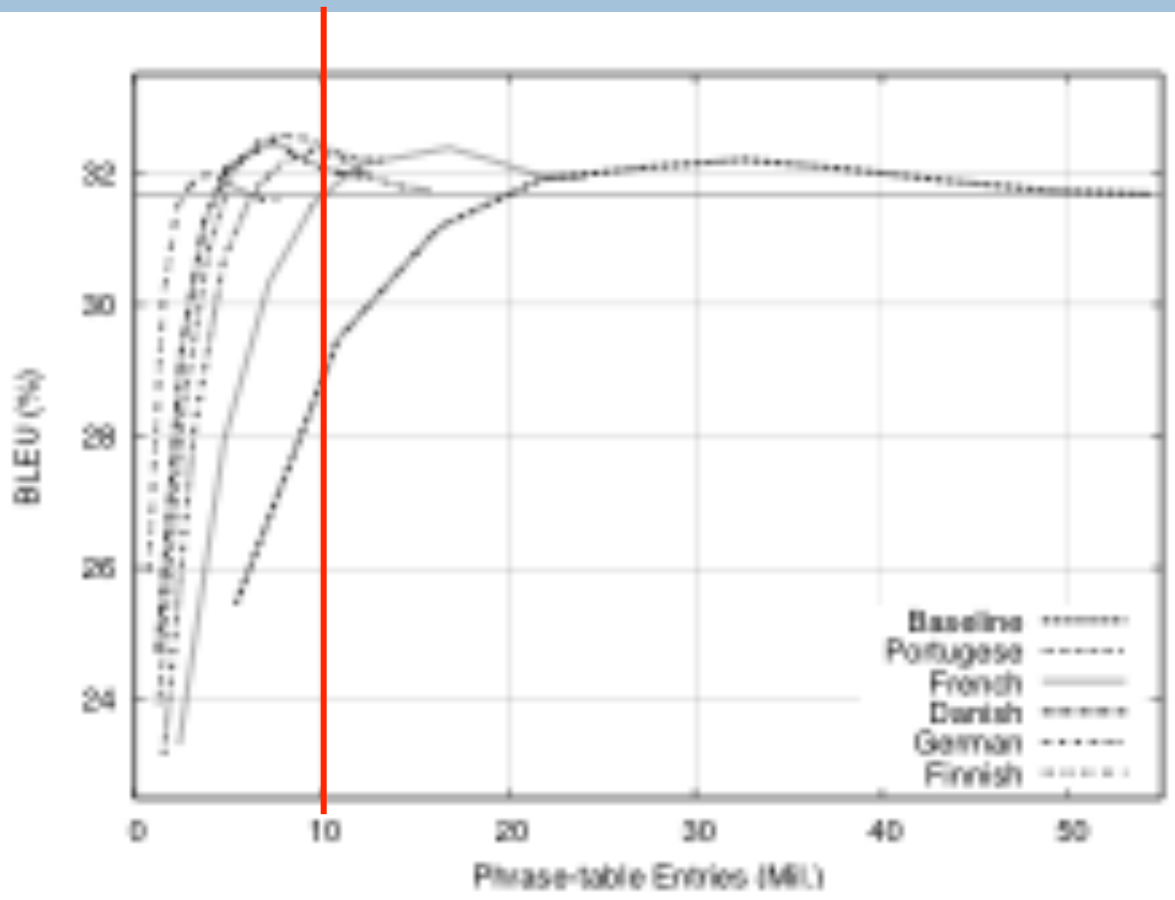
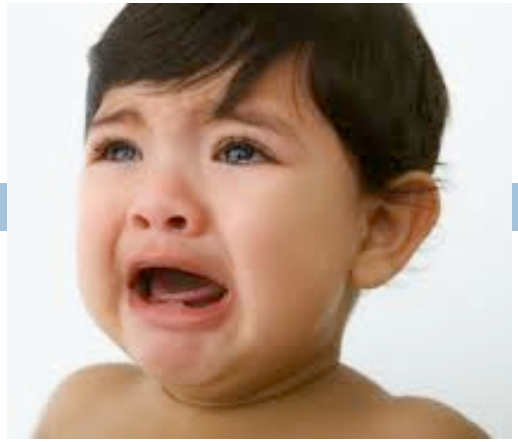


Figure 3: Combining probability-based filtering

Further Research



- German => English

Bridge	EP-40		EP-50	
—	5.1G	26.92	6.5G	27.23
Dutch	562M	27.11	1.3G	28.14
Spanish	3.0G	27.28	3.6G	28.09
Danish	505M	28.04	780M	28.21

Table 7: Filtered German-English systems (Size and BLEU)

Problems / Conclusion

- Throwing away a lot of data (94%)
- Picking a bridge language (sometimes)
- Limitations of data set (Arabic? Chinese?)
- Overall, great!

Thanks!