Inducing Translation Lexicons via Diverse Similarity Measures and Bridge Languages

Charles Schafer and **David Yarowsky**
Department of Computer Science and
Center for Language and Speech Processing
Johns Hopkins University
Baltimore, MD 21218 USA
{cschafer,yarowsky}@cs.jhu.edu

Abstract

This paper presents a method for inducing translation lexicons between two distant languages without the need for either parallel bilingual corpora or a direct bilingual seed dictionary. The algorithm successfully combines temporal occurrence similarity across dates in news corpora, wide and local cross-language context similarity, weighted Levenshtein distance, relative frequency and burstiness similarity measures. These similarity measures are integrated with the bridge language concept under a robust method of classifier combination for both the Slavic and Northern Indian language families.

1 Lexicon Induction via Bridge Languages

The explosive growth of the web in the past 8 years has yielded a corresponding growth in the number of world languages for which text is now available online. Our laboratory alone has been able to acquire O(50,000)-O(100,000,000) words in each of 60 different languages, and increasing quantities of electronic text can readily be found in over 100 world languages. As computers and the internet grow ubiquitous, this trend is extending more and more to small linguistic communities.

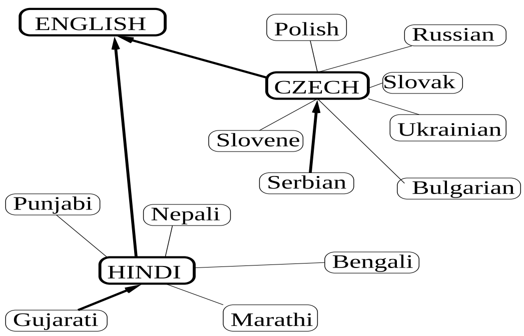

Figure 1: Translingual bridge language pathways for the Slavic and Northern Indian language families.

Universal access to this newly available wealth of information is an important goal, motivating interest in such applications as text glossing, cross-language information retrieval, and ultimately in machine translation. An essential component of each of these applications is the translation lexicon. We would like to develop such a resource between English and each of the languages for which data is available on the internet. Yet from the perspectives of machine translation and machine learning, the problem is daunting, particularly if treated as learning tasks between k independent language pairs. Fortunately, however, languages are not unrelated, and tend to cluster into families and subfamilies with intra-group affinities that can be exploited.

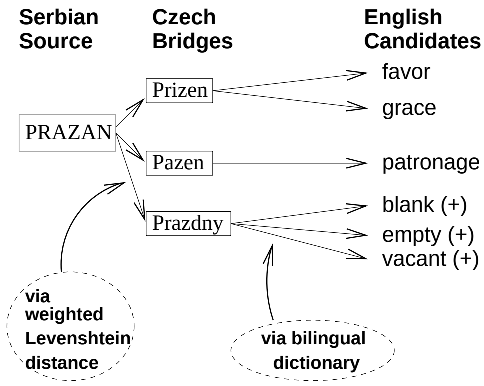

Figure 2: An example Serbian→Czech→English bridge pathway for the Serbian word *Prazan*

Mann and Yarowsky (2001) first proposed the use of these familial affinities for translingual lexicon induction using bridge languages. In particular, they decomposed the process of translation between two languages such as Serbian to English, into a two-step process utilizing another language in the Slavic family such as Czech, for which a sufficiently large and detailed translation lexicon to English is available. Thus: $P(English|Serbian)$

$$= P(English|Czech) \times P(Czech|Serbian)$$

And similarly for any language L_i sufficiently close to Czech:

$$P(English|L_i) = P(English|Czech) \times P(Czech|L_i)$$

Figure 2 illustrates the paradigm of using a bridge language to create a set (ordered by string distance) of candidate English translations for a Serbian word. In Mann and Yarowsky (2001) the process of intra-family translation was handled by weighted string distance models of cognate similarity with a probabilistic representation of common intra-family orthographic transformations. These models were iteratively reestimated using an Expectation-Maximization algorithm (Ristad and Yanilos 1997). When intra-family orthographic shifts are clear and systematic, such models can be quite effective on their own. In practice, the technique described suffers from the problem of *faux amis* – false cognates. For example, Serbian-Czech *faux amis* such as *prazan-prizen* and *prazan-pazen* can outrank the correct but orthographically less similar *prazan-prazdny*, causing the English bridge pathways to the correct English translations *blank* and *empty* to be scored below the incorrect translation paths to *favor*, *grace* and *patronage*.

This paper addresses the above-described model deficiency by proposing, developing and evaluating the use of 7 additional similarity models which successfully capture a set of complementary distributional behaviors. An algorithm combining them with weighted string distance significantly outperforms the previous bridge language approach on both English-Serbian and English-Gujarati test sets.

2 Resources

Our goal was to learn translation lexicons using resources that are available on the internet at no monetary cost. No seed dictionary is required between English and the language of interest; a sizeable dictionary between the bridge language and English is necessary. Our work with Serbian involved the use of a Czech-English dictionary initially containing roughly 171K Czech-English pairs, including 54K unique Czech word types and 43K unique English types. The Hindi-English dictionary contained around 74K pairs. The Serbian/Gujarati vocabularies we used were built by extracting all word types from the respective corpora, then filtering out low-frequency words (since our similarity models require reliable corpus statistics) and very short words[1] (use of string distance to propose cognate candidates for very short words was seen to be unreliable in preliminary experiments). The corpora used here are composed of news data, the majority of which was downloaded from the internet. The English corpus contains 192M tokens; Serbian, 12M; Gujarati, 2M. English was lemmatized using a high-quality lemmatization utility; the Serbian, using minimally supervised morphological analysis as described in Yarowsky and Wicentowski (2000). Gujarati was not lemmatized. Where possible, date labels were extracted for news stories. This resulted in 1690 separate labeled days of news for Serbian and 233 for Gujarati. For each language task, English news data was marked as originating either locally or non-locally with respect to areas where the language is spoken, in order to facilitate computation of date-distributional similarities across both strongly related, same-region news sources (*date-local*) and a general, worldwide aggregate news corpus (*date-all*).

[1] Words with length < 5 characters were excluded.

3 Translation Similarity Models

The algorithm presented here is based on the novel combination of the following 4 categories of similarity models: string similarity, context similarity, date distributional similarity, and similarity of word frequency and burstiness statistics. Three of these 4 categories are further broken down into individual similarity measures for a total of 8: weighted Levenshtein (string), wide and narrow context, world-news and local-news-based date similarities, and relative frequency, burstiness, and inverse document frequency (IDF) similarities. The algorithm used for rank-based combination of the individual models is given in Section 4.

The initial set of candidate translation pairs is generated (as in Figure 2) by considering all source-language words within a low, initially-weighted string distance to entries in the given bridge-language-to-English dictionary. The resulting source-language-to-English candidate pairs are then filtered and ranked by the similarity measures described below:

3.1 Weighted Levenshtein Similarity

On the first iteration, Levenshtein string edit distance uses a simple language-independent matrix that assigns $dist(Vowel^+, Vowel^+)$ and other vowel cluster operations one half the cost of equivalent single consonant substitutions, insertions and deletions.

At the beginning of the 2nd model iteration, the character-distance matrix is reestimated as in Mann and Yarowsky (2001) using the high-confidence output from the 1st iteration as training data. For each of the top 2000 Serbian-English proposed translation pairs after the 1st iteration, the Serbian word and the Czech bridge words having lowest string distance to it (there might have been multiple possible Czech bridges at several distances) are used as a pair into the training set for learning of edit weights.

Some high probability Serbian-Czech orthographic substitutions that are discovered by this process are:

Serbian	Czech	logprob
a	e	-4.6
i	y	-5.8
s	c	-7.2
n	l	-7.5
s	z	-7.7
k	c	-7.7

3.2 Context Similarity

We generate bag-of-words context vectors for both wide (radius 10) and narrow (radius 1) windows surrounding each word in our corpora, for both English and the source language (Serbian, Gujarati). The source language vectors are then translated, using the current iteration's noisy translation lexicon, into English. The initial lexicon is generated by taking the Czech-English bridge dictionary, computing the set of low-edit-distance Serbian-Czech word pairs, and treating the resulting expansion of Serbian-(via Czech)-English word pairs as an initial noisy pair space. Subsequent iterations utilize the translation lexicons induced in the previous training iteration.

This technique is similar to the one presented in Rapp (1999), which uses the concept of cross-language vector similarity to identify English translations of German words. However, under Rapp's method context vectors are translated using roughly 16,000 word pairs from an existing German-English dictionary. Fung (1998) used an approach similar to Rapp's, also starting from a large (20,000 entries) pre-existing bilingual dictionary (Chinese-English). Our approach has the important distinction of utilizing absolutely *no* translation lexicons from the test language to or from any other language, making it suitable for lower density languages.

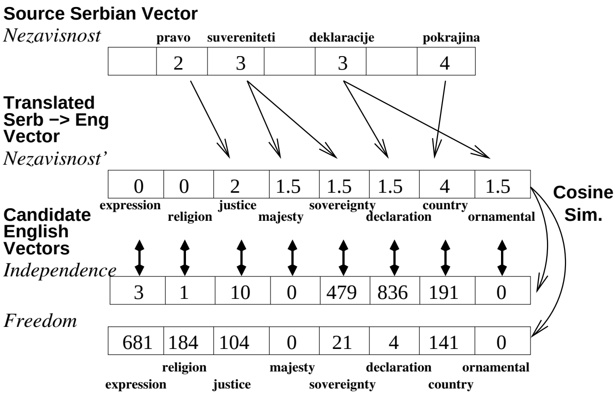

Figure 3: Illustration of the narrow-cosine projection model, comparing Serbian context vectors for the test word *nezavinost* with two candidate English translations based on the previous iteration's translation model. The correct translation of *nezavinost* (*independence*) exhibits greater cosine similarity to the *nezavisnost′* vector than does the competing alternative *freedom*.

3.3 Date-distributional Similarity

One of the advantages of using news data as a corpus is that world and regional events (such as plane crashes, earthquakes, coups, assassinations, etc.) tend to be reported in parallel in multiple languages at reasonably close date synchronization (typically no more than ± 2 days' variation in the reporting of any one story, although followup references persist and degrade over time). Thus both Serbian and English terms can be represented as language-independent frequency vectors subscripted by date over a several-year window. We construct such term vectors for each word in the Serbian and English vocabularies, with frequencies smoothed across adjacent dates to compensate for lags in reportage and to ameliorate sparse data problems. We compiled date distributions for each English word using both worldwide (all English date-labeled news) and local (English news from Serbia) news sources. Given the relatively small size of the local English corpus, incorporated both of these date-distributional models (*date-local* and *date-all*) into our framework for increased robustness.

The example in Figure 4 shows graphically how a (correctly) hypothesized translation pair of *nezavisnost-independence* has greater synchronization in their date distributions, and hence a higher date-similarity score, than a competing incorrect candidate pair of *nezavisnost-freedom*, which has higher-ranked weighted-Levenshtein similarity in Table 5, but ranks lower in the final combined similarities in part due the contribution of date-similarity.

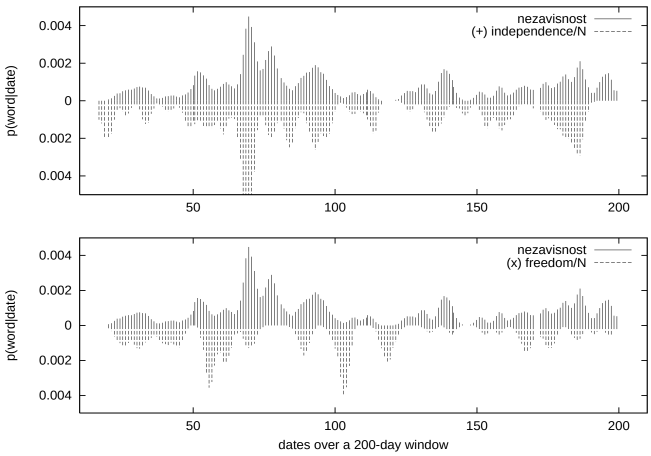

Figure 4: Comparison of the relative date-distributions for the correct translation pair *nezavisnost-independence* (sim=0.74) and the incorrect pair *nezavisnost-freedom* (sim=0.42). In both cases the normalized Serbian word probability is on the positive y-axis, English on the negative y-axis.

3.4 Relative Frequency Similarity

On average, a word and its translation are likely to have similar relative frequencies in the corpora of their respective languages[2]. Because polysemous usage for one language's term may double or triple its observed raw frequency, modest frequency variations are expected. However, this measure is very useful for ruling out hypothesized pairings exhibiting several orders of magnitude difference in relative fre-

[2]Especially when computing frequency similarity on lemmas as is done here. Although individual inflected word frequencies are quite sensitive to the inflectional fertility of the language, the total lemma frequencies for equivalent concepts should be much more consistent across languages.

quency. A simple ratio of logs of frequencies proves to correlate well with translational compatibility and was found to an improvement under the rank-based combination model.

$$RFScore = MIN[\frac{log(rf_1)}{log(rf_2)}, \frac{log(rf_2)}{log(rf_1)}]$$

EW	RF(EW) ($\text{x}10^{-7}$)	RF(hvaliti) ($\text{x}10^{-7}$)	$RFScore_i$
bless/V	64	62	0.998
laud/V	49	62	**0.980**
calibre/N	13	62	0.887
quarter/V	3	62	0.795
class/N	989	62	0.770

Table 1: This table shows the relative frequency (RF) match for the Serbian word *hvaliti*. Its correct translation (in bold) scores higher than alternate hypotheses such as *calibre/N* and *class/N*. Although they outscore *laud/V* on weighted string similarity, their observed 13 and 989 relative frequencies are significantly lower and higher (respectively) than the 62 expected for *hvaliti*'s translation.

3.5 Burstiness Similarity and Inverse Document Frequency

Church and Gale (1995) describe several related measures of a word's tendency towards contagious distributions, such as illustrated in Figure 5. They include the $P_{21}(w)$ measure of adaptability ($P(f_w >= 2 | f_w > 1)$) and standard Inverse Document Frequency (IDF). We used the ratio of IDF's as one of the similarity measures. Given the high variability of document lengths in the corpus, we also defined and utilized a variant measure β over a moving window of H=200 words:

$$\beta = \frac{P(w_i = w | w \in \{w_{i-1}, ..., w_{i-H}\})}{1 - (1 - P(w))^H}$$

$$\beta Match_i = MIN[\frac{\beta_1}{\beta_2}, \frac{\beta_2}{\beta_1}]$$

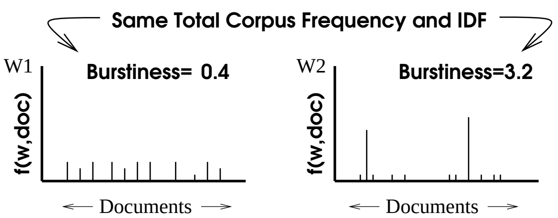

Figure 5: Illustration of the burstiness measure. Both (hypothetical) words have the same total corpus frequency and IDF, yet differ substantially in their burstiness score.

3.6 Use of Additional Bridge Languages

Use of a second bridge language within the source language's family can expand the coverage and/or precision of the bridge dictionary. We investigated

EW	β(EW)	β(hvaliti)	$\beta Match_i$
class/N	2.8	3.0	0.95
laud/V	3.5	3.0	**0.85**
praise/V	2.5	3.0	**0.83**
bless/V	3.9	3.0	0.75
calibre/N	4.9	3.0	0.60
quarter/V	5.1	3.0	0.58
chop/V	5.3	3.0	0.56

Table 2: This table shows the efficacy of the β measure in ranking the correct translation candidates of Serbian *hvaliti* over competitors *calibre/N, quarter/V* and chop/V which have a higher weighted Levenshtein score.

the effect on performance of adding Bulgarian as a second bridge from Serbian to English, which increased accuracy by a relatively consistent by 5-10%.[3]

4 Combining Similarity Measures

Weighted Levenshtein distance initially proposes a set of candidate translation pairs. For each pair in this set, the above-described similarity values are computed. Specifically, the following 8 similarity variants are used: weighted Levenshtein distance (converted to a similarity, i.e., an increasing function of relatedness), wide (radius 10) bag-of-words context similarity, narrow (radius 1) context similarity, local news date distribution similarity, all news date similarity, relative frequency (RF) similarity, inverse document frequency (IDF) similarity, and burstiness (β) similarity.

These individual models are integrated into a single similarity function using the method of rank-based combination. We have observed in previous studies that combining ranks rather than raw scores is more robust and accurate when scores have different dynamic ranges, as they do here. The procedure is as follows for each word s_a in the Serbian vocabulary (For Gujarati, upon which no lemmatization was performed, Step 1 is omitted.):

1. *Part-of-Speech (POS) Consistency:* When ranking translation pairs, we imposed a strong bias in favor of compatible coarse-grained parts of speech (noun, verb, adjective). Each Serbian word is assigned a POS via morphological analysis, and each English translation candidate with a dictionary POS that does not match are given a score penalty sufficient to rank them below POS-compatible candidates, but not exclude them (given possible gaps and errors in POS assignment).

2. *Ranking:* For each similarity measure S, the English candidates are sorted in decreasing order by similarity score. The N English words in this sorted list are assigned counts starting at 0 for the first list item, through $N - 1$ at the last item. Each English word, e_b, having count value c is assigned nor-

[3]Effect on performance is shown in Figure 8 and Table 4.

malized rank $r_{norm}(s_a, e_b, S) = c/N$. Where there are tied similarity values at list positions $i..j$, each tied word is given normalized rank $r_{norm}(s_a, e_b, S) = (c_i + c_j)/N$.

3. *Scoring:* Each similarity model $S_1..S_8$ has an associated weight ($\lambda_1..\lambda_8$) (see Figure 6 for details). For each English word e_b, the rank-based combination score is then computed:

$$R(s_a, e_b) = \sum_{m:1..8} \lambda_m * r_{norm}(s_a, e_b, S_m)$$

Table 5 illustrates the independent performance of the different similarity measures over three Serbian and one Gujarati example test words. Each list is sorted in rank order by descending similarity score.

Additional iterations proceed by retraining the weighted string and narrow/wide context, as described above, using the translation pairs that rank highest on the previous iteration's combined score as initial training data for the next iteration.

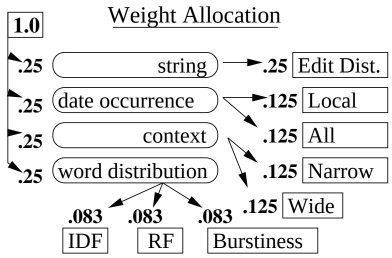

Figure 6: Allocation of weights for ranked-based model combination. As shown, the allocation scheme divides weights up equally per class of similarity model, and performs another equal division of weights for individual models inside a class.

5 Evaluation

Three primary evaluation measures are employed in this study. The first is exact match accuracy of the first choice translation candidate. Another is the percent of cases where a correct English translation is ranked somewhere in the top k hypothesized answers. The third is median position of the per-word highest-ranked correct translation in the system output list. The latter is useful and appropriate because many applications, including cross-language information retrieval and the seeding of alignment lexicons for statistical machine translation, can tolerate some noise in the lexicon, and we want a way to judge how often there is a correct translation close enough to the top of the ranked output to be useful.The tables that follow provide a basis for making such a judgment.

Another issue in evaluation is the accuracy and completeness of the translation lexicon used for scoring. If valid translations are omitted from the gold-standard "truth" set for whatever reason, then systems will be penalized for picking valid answers including a synonym or alternate translation not in the truth list. In an effort to gauge the impact of these truth-set gaps, a second evaluation was performed on a small, randomly selected set of Serbian and Gujarati test words in which a much larger comprehensive hard-copy dictionary (Serbian) or a native speaker (Gujarati) was used to identify additional valid translations. Under this more exhaustive evaluation standard a translation candidate is considered correct if it is either listed in the larger dictionary, is an English synonym of any of the listed truth words, or (for Gujarati) the native speaker judges the words to be synonymous. Serbian exact-match results from both automatic scoring on the full system output and automatic+paper-dictionary scoring on the random subset (scaled to estimate true performance on the full data set) are shown in Figure 7 and Table 3 (also including Gujarati). For many applications, such as generating new candidates for statistical MT alignment and translation models, appearance in an n-best pool of candidates may be as functionally useful as only 1-best exact matches to the limited truth set. This in-n-best accuracy is also given in Table 3.

Finally, Table 4 shows the performance improvement over string distance models yielded by the additional similarity models described in this paper, for the clear case where a correct English translation is known to be in the proposed candidate set (on the basis of the online Serbian-English truth dictionary). As the table shows, for Serbian-English via the Czech bridge, a 9% improvement in exact-match accuracy over the Mann & Yarowsky (2001) trained string distance bridge model is realized in the strongest rank-based combination system. Results of ablation experiments showing the contribution of each class of similarity model are presented in Figure 8.

6 Conclusion

This paper has presented an original technique for inducing translation lexicons via combination of iteratively trained similarity measures. Joint modeling of such a large space of 8 diverse translation-candidate similarity measures is novel and makes effective use of the independence exhibited by several of the evidence sources. In contrast to previous studies such as Fung (1998) and Rapp (1999) that have utilized large (16,000-20,000 word) dictionary subsets as functionally necessary seeds to their context-similarity models, the methods presented here require *no* translation lexicons from the test language to or from any other language. These methods, then, address vocabulary learning for resource limited languages such as Serbian and Gujarati. By taking all necessary supervision from unannotated monolingual texts and 3rd-language dictionary resources, these methods offer great promise for the automatic learning of minority language translation lexicons.

Language	Serbian	Gujarati
Alignment Pool size	(via Czech) 4500 words	(via Hindi) 6200 words
No Bridge available	.26	.35
1+ Bridges available	.74	.65
Accuracy when bridge available (and on full vocabulary):		
Correct in Top 1	.58 (.43)	.46 (.30)
Correct in Top 2	.81 (.60)	.80 (.52)
Correct in Top 3	.85 (.63)	.86 (.56)
Correct in Top 10	.92 (.68)	.89 (.58)

Table 3: A breakdown (by % of test words) that a correct answer appears in the top k of the system's ranked list. Note that in a significant percentage of the time (26-35%) a correct answer is not possible because no valid translation pair had a bridge word within a minimal Levenshtein threshhold to the given (Czech or Hindi) bridge dictionary. This number can be reduced by either augmenting the bridge language dictionary, adding additional bridge languages or both. Evaluation is based on a randomly selected subset scored using exact-match agreement of each hypothesis with translations in the online+paper dictionaries as in Figure 7.

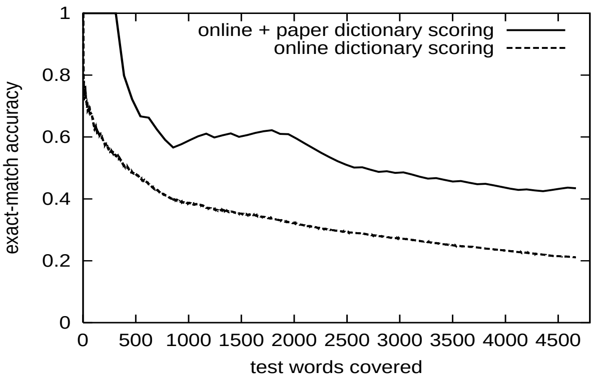

Figure 7: Serbian system performance using both automatic exact-match scoring over an online test dictionary, and manually assisted scoring that also considers an answer correct if it appears in a larger paper dictionary or is a direct synonym of an entry in either dictionary. X-axis is words covered in test vocabulary (sorted by decreasing system confidence, i.e., normalized rank sum of word's top answer). Vocabulary size is roughly 4500 for online-dictionary scoring, and extrapolated from a randomly selected 60-word test sample for manually-assisted online+paper dictionary scoring.

In..	Czech Basic Leven.	Czech Retrained Leven.	Czech Iter 1	Czech Iter 2	Czech +Bulg. Iter 2
Top 1	.36	.39	.47	.48	.51
Top 2	.45	.48	.61	.62	.63
Top 10	.76	.80	.85	.86	.85

Table 4: Serbian performance over all words with a known English translation in the Levenshtein-proposed candidate set, using exact-match to the online Serbian-English evaluation dictionary only. This highlights the gain in performance over both preliminary and retrained string distance models, which is realized by incorporating the full range of similarity components described here (Czech Iter 1) and, additionally, utilizing both context-model retraining (Czech Iter 2) and Bulgarian as a second bridge language (Czech+Bulg Iter 2).

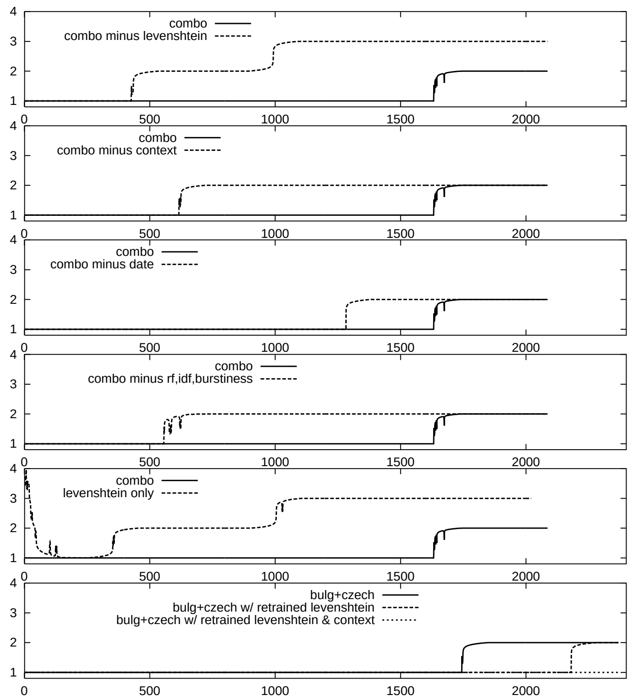

Figure 8: Median rank of correct answer in the confidence-sorted set of translation candidates. This figure shows the relative contributions of the participating similarity measures (string; date-local and date-all; narrow and wide context; burstiness, IDF and relative frequency) by showing the drop in performance given the omission of each category of measures. Clearly string similarity is the single most important model, but the joint performance of the other measures when omitting string similarity exceeds the performance of string similarity alone over the set of translation candidates. To facilitate the relative comparison of methods, this test data set is restricted to the 2300 words which appear in the online scoring dictionary *and* for which at least some bridge path exists between Serbian word and a correct English translation. All median ranks are based on selecting between the candidate English words within a weighted Levenshtein distance of up to 3 from at least one possible bridge term.

References

Church, K. W. and W. A. Gale, 1995. Poisson mixtures. *Natural Language Engineering*, 1(2): pp.163-190.

Fung, P., 1998. In D. Farwell, L. Gerber., and E. Hovy, eds., A Statistical View on bilingual lexicon extraction: From parallel corpora to non-parallel corpora. In *Proceedings of AMTA-98: Machine Translation and the Information Soup*, pp. 1-17. Springer-Verlag.

Koehn, P. and K. Knight, 2001. Estimating Word Translation probabilities from unrelated monolingual corpora using the EM algorithm. In *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pp. 27–35.

otpor (serbian):

RANK	CMB. SCR.	COMBINED	STRING	DATE-LOCAL	WIDE-COS	NARROW-COS	BURSTINESS	RF
1	0.18	protest/N	(1) abhorrence/N	break/V	protest/V	protest/N	protest/N	protest/N
2	0.19	opening/N	(1) abomination/N	**resistance/N**	protest/N	system/N	reluctance/N	port/N
3	0.24	break/N	(1) allergy/N	stress/V	break/V	break/V	break/N	opening/N
4	0.28	mouth/N	(1) animosity/N	protest/V	hate/V	protest/V	kick/V	stress/V
5	0.29	objection/N	(1) antagonism/N	escape/V	opening/N	antagonism/N	protest/V	protest/V
6	0.30	**resistance/N**	(1) antipathy/N	protest/N	escape/V	hate/V	escape/V	escape/V
7	0.30	**opposition/N**	(1) aperture/N	opening/N	stress/V	dislike/V	**opposition/N**	**resistance/N**
8	0.33	reluctance/N	(1) averse/J	break/N	system/N	resentment/N	mouth/N	break/N
9	0.33	port/N	(1) aversion/N	kick/V	defiance/N	unit/N	unit/N	break/V
10	0.36	hole/N	(1) bore/N	system/N	mouth/N	disgust/V	formation/N	**opposition/N**
11	0.38	stress/N	(1) bore/V	**opposition/N**	contradiction/N	reluctance/N	port/N	unit/N
12	0.38	escape/N	(1) boring/J	kick/N	kick/V	formation/N	stress/V	hole/N
13	0.38	formation/N	(1) boring/N	formation/N	resentment/N	animosity/N	objection/N	kick/V
14	0.40	animosity/N	(1) break/N	punch/N	dislike/V	dislike/N	protestation/N	outlet/N
15	0.40	resentment/N	(1) break/V	unit/N	reluctance/N	escape/V	hate/V	column/N
			(1) **opposition/N**		(33) **opposition/N**	(21) **opposition/N**	(20) **resistance/N**	
			(1) **resistance/N**		(53) **resistance/N**	(29) **resistance/N**		

nezavisnost (serbian):

RANK	CMB. SCR.	COMBINED	STRING	DATE-LOCAL	WIDE-COS	NARROW-COS	BURSTINESS	RF
1	0.02	**independence/N**	(1) freedom/N	**independence/N**	**independence/N**	**independence/N**	evidence/V	free/V
2	0.09	freedom/N	(1) **independence/N**	relation/N	single/J	ease/N	necessity/N	cold/J
3	0.11	depend/V	(1) **independence/V**	free/J	cold/N	irrelevant/J	fair/J	abandon/V
4	0.13	relation/N	(4) irrelevance/N	side/N	side/N	ease/V	single/V	importance/N
5	0.20	consequence/N	(5) illegality/N	importance/N	independent/J	applicability/N	application/N	ease/V
6	0.21	lift/V	(5) illegitimacy/N	depend/V	consequence/N	single/J	**independence/N**	licence/N
7	0.21	importance/N	(7) depend/N	independent/J	freedom/N	disagreement/N	currency/N	lift/V
8	0.22	obligation/N	(7) depend/V	single/J	abandon/V	lift/V	free/V	miss/N
9	0.23	ease/V	(7) dependence/N	life/N	lack/V	cold/N	inadequacy/N	green/N
10	0.23	independent/J	(7) dependency/N	freedom/N	depend/V	depend/V	pride/N	involvement/N
11	0.23	single/J	(7) disinterest/J	irrelevant/N	moment/N	pride/N	cold/J	green/J
12	0.24	abandon/V	(7) functionality/N	miss/V	importance/N	side/N	irrelevant/J	consequence/N
13	0.24	integrity/N	(7) innocence/N	imperative/J	relation/N	realty/N	side/V	utility/N
14	0.24	necessity/N	(7) purity/N	safety/N	lack/N	consequence/N	disagreement/N	lack/V
15	0.24	irrelevant/J	(7) relation/N	obligation/N	necessity/N	drag/N	independent/N	independent/N
								(25)**indpndnce/N**

hvaliti (serbian):

RANK	CMB. SCR.	COMBINED	STRING	DATE-LOCAL	WIDE-COS	NARROW-COS	BURSTINESS	RF
1	0.41	**praise/V**	(1) caliber/N	quarter/N	currency/N	currency/N	exchange/V	bless/V
2	0.43	chop/V	(1) calibre/N	good/J	applaud/V	**praise/V**	making/N	chop/V
3	0.45	bless/V	(1) chop/N	quality/N	**praise/N**	superior/J	**praise/N**	commend/V
4	0.48	applaud/V	(1) chop/V	class/N	**praise/V**	good/J	class/N	**laud/V**
5	0.49	exchange/V	(1) class/N	exchange/N	good/J	class/N	currency/N	making/N
6	0.55	**laud/V**	(1) class/V	compliment/N	making/N	good/N	applaud/V	applaud/V
7	0.56	commend/V	(1) making/N	superior/J	bless/V	quarter/N	quarter/N	superior/J
8	0.57	class/V	(1) quality/J	exchange/V	superior/J	quality/N	superior/J	**praise/N**
9	0.68	quarter/V	(1) quality/N	superior/N	good/N	biennial/J	good/N	superior/N
10	0.71	compliment/V	(1) quarter/N	**praise/V**	exchange/V	exchange/N	quality/N	compliment/N
11	0.81	scroll/V	(1) quarter/V	**praise/N**	chop/V	bless/V	superior/N	scroll/N
12	2.30	superior/J	(12) applaud/V	good/N	exchange/N	**praise/N**	**laud/V**	exchange/V
13	2.30	class/N	(12) biennial/J	bless/V	quality/N	exchange/V	**praise/V**	chop/N
14	2.34	quality/N	(12) biennial/N	currency/N	class/N		exchange/N	good/N
15	2.35	making/N	(12) bless/N	caliber/N	biennial/J		bless/V	calibre/N
17	2.41	**praise/N**	(12) **laud/N**	(19) **laud/V**	(18) **laud/V**		(29) **laud/N**	(21) **praise/V**
32	2.83	**laud/N**	(12) **laud/V**		(33) **laud/N**			(25) **laud/N**
			(12) **praise/N**					
			(12) **praise/V**					

uthvayea (gujarati):

RANK	CMB. SCR.	COMBINED	STRING	DATE-LOCAL	WIDE-COS	NARROW-COS	BURSTINESS	RF
1	0.23	**stand/V**	(1) bear/V	rise/V	bear/V	widow/N	**stand/V**	horse/N
2	0.30	suffer/V	(1) endure/V	suffer/V	**stand/V**	**stand/V**	raise/V	expire/V
3	0.31	bear/V	(1) expire/V	**stand/V**	leave/V	leave/V	suffer/V	proceed/V
4	0.39	leave/V	(1) leave/V	limit/N	suffer/V	bear/V	bear/V	quantity/N
5	0.41	proceed/V	(1) proceed/V	raise/V	endure/V	boundary/N	leave/V	boundary/N
6	0.41	endure/V	(1) raise/V	bear/V	limit/N	endure/V	rise/V	limit/N
7	0.42	raise/V	(1) rise/V	leave/V	raise/V	limit/N	proceed/V	endure/V
8	0.44	rise/V	(1) shallow/J	horse/N	quantity/N	suffer/V	endure/V	widow/N
9	0.45	expire/V	(1) **stand/V**	boundary/N	proceed/V	proceed/V	limit/N	bear/V
10	0.45	limit/N	(1) suffer/V	expire/V	horse/N	raise/V	expire/V	suffer/V
11	0.52	boundary/N	(11) mischief/N	quantity/N	widow/N	expire/V	quantity/N	**stand/V**
12	0.57	quantity/N	(12) boundary/N	proceed/V	boundary/N	rise/V	widow/N	mischief/N
13	0.61	widow/N	(12) horse/N	endure/V	shallow/J	horse/N	horse/N	raise/V
14	0.62	horse/N	(12) limit/N	widow/N	rise/V	quantity/N	boundary/N	shallow/J
15	0.72	shallow/J	(12) quantity/N	mischief/N	expire/V		shallow/J	rise/V

Table 5: These tables show the performance of individual similarity measures as well as their combined choice, after model retraining. Correct translations are shown in bold. Note that in many cases the string-similarity-based orderings of the bridge candidates underperform individual non-string similarity measures, and they consistently underperform the weighted combination of all 8 similarity measures. Note that in the case of *nezavisnost*, the correct translation *independence* is successfully ranked above its quite closely related competitor *freedom* by almost every non-string-based similarity measure in isolation. This behavior (shown quantatively in Figure 8) illustrates the contribution of consensus modeling over this set of diverse similarity measures.

Mann, G. and D. Yarowsky, 2001. Multipath translation lexicon induction via bridge languages. In *Proceedings of NAACL-2001*, pp. 151–158.

Rapp, R., 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of ACL-99*, pp. 519–526.

Ristad, E. S., and P. N. Yianilos, 1997. Learning string edit distance. In *Machine Learning: Proceedings of the Fourteenth International Conference*, pp. 287–295.

Yarowsky, D. and R. Wicentowski, 2000. Minimally supervised morphological analysis by multimodal alignment. In *Proceedings of ACL-2000*, pp. 207–216.