Modelling lexical redundancy for machine translation

David Talbot and Miles Osborne
School of Informatics, University of Edinburgh
2 Buccleuch Place, Edinburgh, EH8 9LW, UK
d.r.talbot@sms.ed.ac.uk, miles@inf.ed.ac.uk

Abstract

Certain distinctions made in the lexicon of one language may be *redundant* when translating into another language. We quantify redundancy among source types by the similarity of their distributions over target types. We propose a language-independent framework for minimising *lexical redundancy* that can be optimised directly from parallel text. Optimisation of the source lexicon for a given target language is viewed as model selection over a set of cluster-based translation models.

Redundant distinctions between types may exhibit monolingual regularities, for example, inflexion patterns. We define a prior over model structure using a Markov random field and learn features over sets of monolingual types that are predictive of bilingual redundancy. The prior makes model selection more robust without the need for language-specific assumptions regarding redundancy. Using these models in a phrase-based SMT system, we show significant improvements in translation quality for certain language pairs.

1 Introduction

Data-driven machine translation (MT) relies on models that can be efficiently estimated from parallel text. Token-level independence assumptions based on word-alignments can be used to decompose parallel corpora into manageable units for parameter estimation. However, if training data is scarce or language pairs encode significantly different information in the lexicon, such as Czech and English, additional independence assumptions may assist the model estimation process.

Standard statistical translation models use separate parameters for each pair of source and target types. In these models, distinctions in either lexicon that are redundant to the translation process will result in unwarranted model complexity and make parameter estimation from limited parallel data more difficult. A natural way to eliminate such *lexical redundancy* is to group types into homogeneous clusters that do not differ significantly in their distributions over types in the other language. Cluster-based translation models capture the corresponding independence assumptions.

Previous work on bilingual clustering has focused on coarse partitions of the lexicon that resemble automatically induced part-of-speech classes. These were used to model generic word-alignment patterns such as noun-adjective re-ordering between English and French (Och, 1998). In contrast, we induce fine-grained partitions of the lexicon, conceptually closer to automatic lemmatisation, optimised specifically to assign translation probabilities. Unlike lemmatisation or stemming, our method specifically quantifies lexical redundancy in a bilingual setting and does not make language-specific assumptions.

We tackle the problem of redundancy in the translation lexicon via Bayesian model selection over a set of cluster-based translation models. We search for the model, defined by a clustering of the source lexicon, that maximises the *marginal likelihood* of target tokens in parallel data. In this optimisation, source types are combined into clusters if their distributions over target types are too similar to warrant distinct parameters.

Redundant distinctions between types may exhibit regularities within a language, for instance, inflexion patterns. These can be used to guide model selection. Here we show that the inclusion of a model 'prior' over the lexicon structure leads to more robust translation models. Although *a priori* we do not know which monolingual features characterise redundancy for a given language pair, by defining a model over the prior monolingual

Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL, pages 969–976,
Sydney, July 2006.

space of source types and cluster assignments, we can introduce an inductive bias that allows clustering decisions in different parts of the lexicon to influence one another via *monolingual* features. We use an EM-type algorithm to learn weights for a Markov random field parameterisation of this prior over lexicon structure.

We obtain significant improvements in translation quality as measured by BLEU, incorporating these optimised model within a phrase-based SMT system for three different language pairs. The MRF prior improves the results and picks up features that appear to agree with linguistic intuitions of redundancy for the language pairs considered.

2 Lexical redundancy between languages

In statistical MT, the source and target lexicons are usually defined as the sets of distinct types observed in the parallel training corpus for each language. Such models may not be optimal for certain language pairs and training regimes.

A word-level statistical translation model approximates the probability $Pr(E|F)$ that a source type indexed by F will be translated as a target type indexed by E. Standard models, e.g. Brown et al. (1993), consist of discrete probability distributions with separate parameters for each unique pairing of a source and target types; no attempt is made to leverage structure within the event spaces $\mathcal{E}$ and $\mathcal{F}$ during parameter estimation. This results in a large number of parameters that must be estimated from limited amounts of parallel corpora.

We refer to distinctions made between lexical types in one language that do not result in different distributions over types in the other language as *lexically redundant* for the language pair. Since the role of the translation model is to determine a distribution over target types given a source type, when the corresponding target distributions do not vary significantly over a set of source types, the model gains nothing by maintaining a distinct set of parameters for each member of this set.

Lexical redundancy may arise when languages differ in the specificity with which they refer to the same concepts. For instance, colours of the spectrum may be partitioned differently (e.g. *blue* in English v.s. *sinii* and *goluboi* in Russian). It will also arise when languages explicitly encode different information in the lexicon. For example, translating from French to English, a standard model would treat the following pairs of source and target types as distinct events with entirely unrelated parameters: $(vert, green)$, $(verte, green)$, $(verts, green)$ and $(vertes, green)$. Here the French types differ only in their final suffixes due to adjectival agreement. Since there is no equivalent mechanism in English, these distinctions are redundant with respect to this target language.

Distinctions that are redundant in the source lexicon when translating into one language may, however, be significant when translating into another. For instance, the French adjectival number agreement (the addition of an *s*) may be significant when translating to Russian which also marks adjectives for number (the inflexion to *-ye*).

We can remove redundancy from the translation model by conflating redundant types, e.g. $\overline{vert} \doteq \{vert, verte, verts, vertes\}$, and averaging bilingual statistics associated with these events.

3 Eliminating redundancy in the model

Redundancy in the translation model can be viewed as unwarranted model complexity. A cluster-based translation model defined via a hard-clustering of the lexicon can reduce this complexity by introducing additional independence assumptions: given the source cluster label, c_j, the target type, e_i, is assumed to be independent of the exact source type, f_j, observed, i.e., $p(e_i|f_j) \approx p(e_i|c_j)$. Optimising the model for lexical redundancy can be viewed as model selection over a set of such cluster-based translation models.

We formulate model search as a *maximum a posteriori* optimisation: the data-dependent term, $p(D|C)$, quantifies evidence provided for a model, C, by bilingual training data, D, while the prior, $p(C)$, can assert a preference for a particular model structure (clustering of the source lexicon) on the basis of monolingual features. Both terms have parameters that are estimated from data. Formally, we search for C^*,

$$\begin{aligned} C^* &= \arg\max_C \; p(C|D) \\ &= \arg\max_C \; p(C)p(D|C). \end{aligned} \quad (1)$$

Evaluating the data-dependent term, $p(D|C)$, for different partitions of the source lexicon, we can compare how well different models predict the target tokens aligned in a parallel corpus. This term will prefer models that group together source types with similar distributions over target types. By using the *marginal likelihood* (integrating out the parameters of the translation model) to calculate

$p(D|C)$, we can account explicitly for the complexity of the translation model and compare models with different *numbers* of clusters as well as different assignments of types to clusters.

In addition to an implicit uniform prior over cluster labels as in k-means clustering (e.g. Chou (1991)), we also consider a Markov random field (MRF) parameterisation of the $p(C)$ term to capture monolingual regularities in the lexicon. The MRF induces dependencies between clustering decisions in different parts of the lexicon via a monolingual feature space biasing the search towards models that exhibit monolingual regularities. Rather than assuming *a priori* knowledge of redundant distinctions in the source language, we use an EM algorithm to update parameters for features defined over sets of source types on the basis of existing cluster assignments. While initially the model search will be guided only by information from the bilingual statistics in $p(D|C)$, monolingual regularities in the lexicon, such as inflexion patterns, may gradually be propagated through the model as $p(C)$ becomes informative. Our experiments suggest that the MRF prior enables more robust model selection.

As stated, the model selection procedure accounts for redundancy in the source lexicon using the target distributions. The target lexicon can be optimised analogously. Clustering target types allows the implementation of independence assumptions asserting that the exact specification of a target type is independent of the source type given knowledge of the target cluster label. For example, when translating an English adjective into French it may be more efficient to use the translation model to specify only that the translation lies within a certain set of French adjectives, corresponding to a single lemma, and have the language model select the exact form. Our experiments suggest that it can be useful to account for redundancy in both languages in this way; this can be incorporated simply within our optimisation procedure.

In Section 3.1 we describe the bilingual marginal likelihood, $p(D|C)$, clustering procedure; in Section 3.2 we introduce the MRF parameterisation of the prior, $p(C)$, over model structure; and in Section 3.3, we describe algorithmic approximations.

3.1 Bilingual model selection

Assume we are optimising the source lexicon (the target lexicon is optimised analogously). A clustering of the lexicon is a unique mapping $C_F : \mathcal{F} \rightarrow \mathcal{C}_F$ defined for all $f \in \mathcal{F}$ where, in addition to all source types observed in the parallel training corpus, $\mathcal{F}$ may include items seen in other monolingual corpora (and, in the case of the source lexicon only, the development and test data). The standard SMT lexicon can be viewed as a clustering with each type observed in the parallel training corpus assigned to a distinct cluster and all other types assigned to a single 'unknown word' cluster.

We optimise a conditional model of target tokens from word-aligned parallel corpora, $D = \{D_{c_0}, ..., D_{c_N}\}$, where D_{c_i} represents the set of target words that were aligned to the set of source types in cluster c_i. We assume that each target token in the corpus is generated conditionally *i.i.d.* given the cluster label of the source type to which it is aligned. Sufficient statistics for this model consist of co-occurrence counts of source and target types summed across each source cluster,

$$\#_{c_f}(e) \doteq \sum_{f' \in c_f} \#(e, f'). \quad (2)$$

Maximising the likelihood of the data under this model would require us to specify the number of clusters (the size of the lexicon) in advance. Instead we place a Dirichlet prior parameterised by α[1] over the translation model parameters of each cluster, $\mu_{c_f,e}$, defining the conditional distributions over target types. Given a clustering, the Dirichlet prior, and independent parameters, the distribution over data and parameters factorises,

$$p(D, \mu|C_F, \alpha) = \prod_{c_f \in \mathcal{C}_F} p(D_{c_f}, \mu_{c_f}|c_f, \alpha)$$
$$\propto \prod_{c_f \in \mathcal{C}_F} \prod_{e \in \mathcal{E}} \mu_{c_f,e}^{\alpha - 1 + \#_{c_f}(e)}$$

We optimise cluster assignments with respect to the *marginal likelihood* which averages the likelihood of the set of counts assigned to a cluster, D_{c_f}, under the current model over the prior,

$$p(D_{c_f}|\alpha, c_f) = \int p(\mu_{c_f}|\alpha) p(D_{c_f}|\mu_{c_f}, c_f) d\mu_{c_f}.$$

This can be evaluated analytically for a Dirichlet prior with multinomial parameters.

Assuming a (fixed) uniform prior over model structure, $p(C)$, model selection involves iteratively re-assigning source types to clusters such as to maximise the marginal likelihood. Reassignments may alter the total number of clusters

[1] Distinct from the prior over model structure, $p(C)$.

at any point. Updates can be calculated locally, for instance, given the sets of target tokens D_{c_i} and D_{c_j} aligned to source types currently in clusters c_i and c_j, the change in log marginal likelihood if clusters c_i and c_j are merged into cluster $\bar{c}$ is,

$$\Delta_{c_i,c_j\to\bar{c}} = \log \frac{p(D_{\bar{c}}|\alpha,\bar{c})}{p(D_{c_i}|\alpha,c_i)p(D_{c_j}|\alpha,c_j)}, \quad (3)$$

which is a *Bayes factor* in favour of the hypothesis that D_{c_i} and D_{c_j} were sampled from the same distribution (Wolpert, 1995). Unlike its equivalent in maximum likelihood clustering, Eq.(3) may assume positive values favouring a smaller number of clusters when the data does not support a more complex hypothesis. The more complex model, with c_i and c_j modelled separately, is penalised for being able to model a wider range of data sets.

The hyperparameter, α, is tied across clusters and taken to be proportional to the marginal (the 'background') distribution over target types in the corpus. Under this prior, source types aligned to the same target types, will be clustered together more readily if these target types are less frequent in the corpus as a whole.

3.2 Markov random field model prior

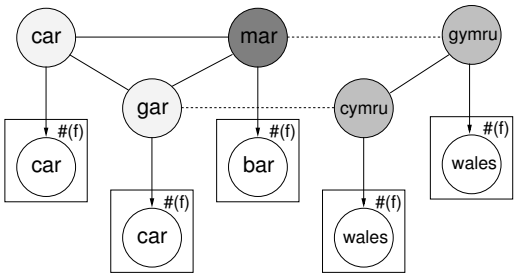
As described above we consider a Markov random field (MRF) parameterisation of the prior over model structure, $p(C)$. This defines a distribution over cluster assignments of the source lexicon as a whole based solely on monolingual characteristics of the lexical types and the relations between their respective cluster assignments.

Viewed as graph, each variable in the MRF is modelled as conditionally independent of all other variables given the values of its neighbours (the Markov property; (Geman and Geman, 1984)). Each variable in the MRF prior corresponds to a lexical source type and its cluster assignment. Fig. 1 shows a section of the complete model including the MRF prior for a Welsh source lexicon; shading denotes cluster assignments and English target tokens are shown as directed nodes.[2] From the Markov property it follows that this prior decomposes over neighbourhoods,

$$p_{\mathbf{MRF}}(C) \propto e^{\beta \sum_{f\in\mathcal{F}} \sum_{f'\in\mathcal{N}_f} \sum_i \lambda_i \psi_i(f,f',c_f,c'_f)}$$

Here $\mathcal{N}_f$ is the set of neighbours of source type f; i indexes a set of functions $\psi_i(\cdot)$ that pick out features of a clique; each function has a parameter λ_i that we learn from the data; these are tied across the graph. β is a free parameter used to control the overall contribution of the prior in Eq. (1). Here features are defined over pairs of types but higher-order interactions can also be modelled. We only consider 'positive' prior knowledge that is indicative of redundancy among source types. Hence all features are non-zero only when their arguments are assigned to the same cluster.

Figure 1: Model with Markov random field prior



Features can be defined over any aspects of the lexicon; in our experiments we use binary features over constrained string edits between types. The following feature would be 1, for instance, if the Welsh types *cymru* and *gymru* (see Fig. 1), were assigned to the same cluster.[3]

$$\psi_1(f_i = (c\sim) \wedge f_j = (g\sim) \ \wedge \ \ c_i = c_j)$$

Setting the parameters of the MRF prior over this feature space by hand would require *a priori* knowledge of redundancies for the language pair. In the absence of such knowledge, we use an iterative EM algorithm to update the parameters on the basis of the previous solution to the bilingual clustering procedure. EM parameter estimation forces the cluster assignments of the MRF prior to agree with those obtained on the basis of bilingual data using monolingual features alone. Since features are tied across the MRF, patterns that characterise redundant relations between types will be re-enforced across the model. For instance (see Fig. 1), if $cymru$ and $gymru$ are clustered together, the parameter for feature ψ_1, shown above, may increase. This induces a prior preference for car and gar to form a cluster on subsequent iterations. A similar feature defined for mar and gar in the *a priori* string edit feature space, on the other hand, may remain uninformative if not observed frequently on pairs of types assigned to the same clusters. In this way, the model learns to

[2]The *plates* represent repeated sampling; each Welsh source type may be aligned to multiple English tokens.

[3]Here $\sim$ matches a common substring of both arguments.

generalise language-specific redundancy patterns from a large *a priori* feature space. Changes in the prior due to re-assignments can be calculated locally and combined with the marginal likelihood.

3.3 Algorithmic approximations

The model selection procedure is an EM algorithm. Each source type is initially assigned to its own cluster and the MRF parameters, λ_i, are initialised to zero. A greedy E-step iteratively reassigns each source type to the cluster that maximises Eq. (1); cluster statistics are updated after any re-assignment. To reduce computation, we only consider re-assignments that would cause at least one (non-zero) feature in the MRF to fire, or to clusters containing types sharing target word-alignments with the current type; types may also be re-assigned to a cluster of their own at any iteration. When clustering both languages simultaneously, we average 'target' statistics over the number of events in each 'target' cluster in Eq. (2).

We re-estimate the MRF parameters after each pass through the vocabulary. These are updated according to MLE using a *pseudolikelihood approximation* (Besag, 1986). Since MRF parameters can only be non-zero for features observed on types clustered together during an E-step, we use lazy instantiation to work with a large implicit feature set defined by a constrained string edit.

The algorithm has two free parameters: α determining the strength of the Dirichlet prior used in the marginal likelihood, $p(D|C)$, and β which determines the contribution of $p_{\mathbf{MRF}}(C)$ to Eq. (1).

4 Experiments

Phrase-based SMT systems have been shown to outperform word-based approaches (Koehn et al., 2003). We evaluate the effects of lexicon model selection on translation quality by considering two applications within a phrase-based SMT system.

4.1 Applications to phrase-based SMT

A phrase-based translation model can be estimated in two stages: first a parallel corpus is aligned at the word-level and then phrase pairs are extracted (Koehn et al., 2003). Aligning tokens in parallel sentences using the IBM Models (Brown et al., 1993), (Och and Ney, 2003) may require less information than full-blown translation since the task is constrained by the source and target tokens present in each sentence pair. In the phrase-level translation table, however, the model must assign probabilities to a potentially unconstrained set of target phrases. We anticipate the optimal model sizes to be different for these two tasks.

Source	Tokens	Types	Singletons	Test	OOV
Czech	468K	54K	29K	6K	469
French	5682K	53K	19K	16K	112
Welsh	4578K	46K	18K	15K	64

Table 1: Parallel corpora used in the experiments.

We can incorporate an optimised lexicon at the word-alignment stage by mapping tokens in the training corpus to their cluster labels. The mapping will not change the number of tokens in a sentence, hence the word-alignments can be associated with the original corpus (see Exp. 1).

To extrapolate a mapping over phrases from our type-level models we can map each type within a phrase to its corresponding cluster label. This, however, results in a large number of distinct phrases being collapsed down to a single 'clustered phrase'. Using these directly may spread probability mass too widely. Instead we use them to smooth the phrase translation model (see Exp. 2). Here we consider a simple interpolation scheme; they could also be used within a backoff model (Yang and Kirchhoff, 2006).

4.2 Experimental set-up

The system we use is described in (Koehn, 2004). The phrase-based translation model includes phrase-level and lexical weightings in both directions. We use the decoder's default behaviour for unknown words copying them verbatim to the output. Smoothed trigram language models are estimated on training sections of the parallel corpus.

We used the parallel sections of the Prague Treebank (Cmejrek et al., 2004), French and English sections of the Europarl corpus (Koehn, 2005) and parallel text from the Welsh Assembly[4] (see Table1). The source languages, Czech, French and Welsh, were chosen on the basis that they may exhibit different degrees of redundancy with respect to English and that they differ morphologically. Only the Czech corpus has explicit morphological annotation.

4.3 Models

All models used in the experiments are defined as mappings of the source and target vocabularies. The target vocabulary includes all distinct types

[4]This Welsh-English parallel text is in the public domain. Contact the first author for details.

seen in the training corpus; the source vocabulary also includes types seen only in development and test data. Free parameters were set to maximize our evaluation metric, BLEU, on development data. The results are reported on the test sets (see Table 1). The baseline mappings used were:

- *standard*: the identity mapping;
- *max-pref*: a prefix of no more than n letters;
- *min-freq*: a prefix with a frequency of at least n in the parallel training corpus.
- *lemmatize*: morphological lemmas (Czech)

standard corresponds to the standard SMT lexicon. *max-pref* and *min-freq* are both simple stemming algorithms that can be applied to raw text. These mappings result in models defined over fewer distinct events that will have higher frequencies; *min-freq* optimises the latter directly. We optimise over (possibly different) values of n for source and target languages. The *lemmatize* mapping which maps types to their lemmas was only applicable to the Czech corpus.

The optimised lexicon models define mappings directly via their clusterings of the vocabulary. We consider the following four models:

- *src*: clustered source lexicon;
- *src+mrf*: as *src* with MRF prior;
- *src+trg*: clustered source and target lexicons;
- *src+trg+mrf*: as *src+trg* with MRF priors.

In each case we optimise over α (a single value for both languages) and, when using the MRF prior, over β (a single value for both languages).

4.4 Experiments

The two sets of experiments evaluate the baseline models and optimised lexicon models during word-alignment and phrase-level translation model estimation respectively.

- Exp. 1: map the parallel corpus, perform word-alignment; estimate the phrase translation model using the original corpus.
- Exp. 2: smooth the phrase translation model,

$$p(\mathbf{e}|\mathbf{f}) = \frac{\#(\mathbf{e},\mathbf{f}) + \gamma\#(\mathbf{c_e},\mathbf{c_f})}{\#(\mathbf{f}) + \gamma\#(\mathbf{c_f})}$$

  Here $\mathbf{e}$, $\mathbf{f}$ and $\mathbf{c_e}$, $\mathbf{c_f}$ are phrases mapped under the *standard* model and the model being tested respectively; γ is set once for all experiments on development data. Word-alignments were generated using the optimal *max-pref* mapping for each training set.

5 Results

Table 2 shows the changes in BLEU when we incorporate the lexicon mappings during the word-alignment process. The standard SMT lexicon model is not optimal, as measured by BLEU, for any of the languages or training set sizes considered. Increases over this baseline, however, diminish with more training data. For both Czech and Welsh, the explicit model selection procedure that we have proposed results in better translations than all of the baseline models when the MRF prior is used; again these increases diminish with larger training sets. We note that the stemming baseline models appear to be more effective for Czech than for Welsh. The impact of the MRF prior is also greater for smaller training sets.

Table 3 shows the results of using these models to smooth the phrase translation table.[5] With the exception of Czech, the improvements are smaller than for Exp 1. For all source languages and models we found that it was optimal to leave the target lexicon unmapped when smoothing the phrase translation model.

Using *lemmatize* for word-alignment on the Czech corpus gave BLEU scores of 32.71 and 37.21 for the 10K and 21K training sets respectively; used to smooth the phrase translation model it gave scores of 33.96 and 37.18.

5.1 Discussion

Model selection had the largest impact for smaller data sets suggesting that the complexity of the standard model is most excessive in sparse data conditions. The larger improvements seen for Czech and Welsh suggest that these languages encode more redundant information in the lexicon with respect to English. Potential sources could be grammatical case markings (Czech) and mutation patterns (Welsh). The impact of the MRF prior for smaller data sets suggests it overcomes sparsity in the bilingual statistics during model selection.

The location of redundancies, in the form of case markings, at the *ends* of words in Czech as assumed by the stemming algorithms may explain why these performed better on this language than

[5] The *standard* model in Exp. 2 is equivalent to the optimised *max-pref* in Exp. 1.

Table 2: BLEU scores with optimised lexicon applied during word-alignment (Exp. 1)

	Czech-English		French-English				Welsh-English			
Model	10K sent.	21K	10K	25K	100K	250K	10K	25K	100K	250K
standard	32.31	36.17	20.76	23.17	26.61	27.63	35.45	39.92	45.02	46.47
max-pref	34.18	37.34	21.63	23.94	26.45	28.25	35.88	41.03	44.82	46.11
min-freq	33.95	36.98	21.22	23.77	26.74	27.98	36.23	40.65	45.38	46.35
src	33.95	37.27	21.43	**24.42**	**26.99**	27.82	36.98	40.98	45.81	46.45
src+mrf	33.97	37.89	21.63	24.38	26.74	**28.39**	37.36	41.13	**46.50**	**46.56**
src+trg	34.24	38.28	22.05	24.02	26.53	27.80	36.83	41.31	45.22	46.51
src+trg+mrf	**34.70**	**38.44**	**22.33**	23.95	26.69	27.75	**37.56**	**42.19**	45.18	46.48

Table 3: BLEU scores with optimised lexicon used to smooth phrase-based translation model (Exp. 2)

	Czech-English		French-English				Welsh-English			
Model	10K sent.	21K	10K	25K	100K	250K	10K	25K	100K	250K
(standard)[5]	34.18	37.34	21.63	23.94	26.45	28.25	35.88	41.03	44.82	46.11
max-pref	35.63	38.81	**22.49**	24.10	**26.99**	28.26	37.31	40.09	45.57	46.41
min-freq	34.65	37.75	21.14	23.41	26.29	27.47	36.40	40.84	45.75	46.45
src	34.38	37.98	21.28	**24.17**	26.88	**28.35**	36.94	39.99	45.75	**46.65**
src+mrf	**36.24**	**39.70**	22.02	24.10	26.82	28.09	**37.81**	**41.04**	**46.16**	46.51

Table 4: System output (Welsh 25K; Exp. 2)

Src	ehangu o ffilm i deledu.
Ref	an expansion from film into television.
standard	expansion of footage to *deledu*.
max-pref	expansion of *ffilm* to television.
src+mrf	expansion of film to television.
Src	yw gwarchod cymru fel gwlad brydferth
Ref	safeguarding wales as a picturesque country
standard	protection of wales as a country *brydferth*
max-pref	protection of wales as a country *brydferth*
src+mrf	protecting wales as a beautiful country
Src	cynhyrchu canlyniadau llai na pherffaith
Ref	produces results that are less than perfect
standard	produce results less than *pherffaith*
max-pref	produce results less than *pherffaith*
src+mrf	generates less than perfect results
Src	y dynodiad o graidd y broblem
Ref	the identification of the nub of the problem
standard	the *dynodiad* of the heart of the problem
max-pref	the *dynodiad* of the heart of the problem
src+mrf	the identified crux of the problem

Table 5: Features learned by MRF prior

Czech	French	Welsh
$(\sim, \sim m)$	$(\sim, \sim s)$	$(c \sim, g \sim)$
$(\sim, \sim u)$	$(\sim, \sim e)$	$(d \sim, dd \sim)$
$(\sim, \sim a)$	$(\sim, \sim es)$	$(d \sim, t \sim)$
$(\sim, \sim ch)$	$(\sim e, \sim es)$	$(b \sim, p \sim)$
$(\sim, \sim ho)$	$(\sim e, \sim er)$	$(c \sim, ch \sim)$
$(\sim a, \sim u)$	$(\sim e, \sim ent)$	$(b \sim, f \sim)$

Note: Features defined over pairs of source types assigned to the same cluster; here $\sim$ matches a common substring.

Table 6: Optimal lexicon size (ratio of raw vocab.)

	Czech	French	Welsh
Word-alignment	0.26	0.22	0.24
TM smoothing	0.28	0.38	0.51

on Welsh. The highest scoring features in the MRF (see Table 5) show that Welsh redundancies, on the other hand, are primarily between *initial* characters. Inspection of system output confirms that OOV types could be mapped to known Welsh words with the MRF prior but not via stemming (see Table 4). For each language pair the MRF learned features that capture intuitively redundant patterns: adjectival endings for French, case markings for Czech, and mutation patterns for Welsh.

The greater improvements in Exp. 1 were mirrored by higher compression rates for these lexicons (see Table. 6) supporting the conjecture that word-alignment requires less information than full-blown translation. The results of the *lemmatize* model on Czech show the model selection procedure improving on a simple supervised baseline.

6 Related Work

Previous work on automatic bilingual word clustering has been motivated somewhat differently and not made use of cluster-based models to assign translation probabilities directly (Wang et al., 1996), (Och, 1998). There is, however, a large body of work using morphological analysis to define cluster-based translation models similar to ours but in a supervised manner (Zens and Ney, 2004), (Niessen and Ney, 2004). These approaches have used morphological annotation (e.g. lemmas and part of speech tags) to provide explicit supervision. They have also involved manually specifying which morphological distinc-

tions are redundant (Goldwater and McClosky, 2005). In contrast, we attempt to learn both equivalence classes and redundant relations automatically. Our experiments with orthographic features suggest that some morphological redundancies can be acquired in an unsupervised fashion.

The marginal likelihood hard-clustering algorithm that we propose here for translation model selection can be viewed as a Bayesian k-means algorithm and is an application of Bayesian model selection techniques, e.g., (Wolpert, 1995). The Markov random field prior over model structure extends the fixed uniform prior over clusters implicit in k-means clustering and is common in computer vision (Geman and Geman, 1984). Recently Basu et al. (2004) used an MRF to embody hard constraints within semi-supervised clustering. In contrast, we use an iterative EM algorithm to *learn* soft constraints within the 'prior' monolingual space based on the results of clustering with bilingual statistics.

7 Conclusions and Future Work

We proposed a framework for modelling lexical redundancy in machine translation and tackled optimisation of the lexicon via Bayesian model selection over a set of cluster-based translation models. We showed improvements in translation quality incorporating these models within a phrase-based SMT sytem. Additional gains resulted from the inclusion of an MRF prior over model structure. We demonstrated that this prior could be used to learn weights for monolingual features that characterise bilingual redundancy. Preliminary experiments defining MRF features over morphological annotation suggest this model can also identify redundant distinctions categorised linguistically (for instance, that *morphological case* is redundant on Czech nouns and adjectives with respect to English, while *number* is redundant only on adjectives). In future work we will investigate the use of linguistic resources to define feature sets for the MRF prior. Lexical redundancy would ideally be addressed in the context of phrases, however, computation and statistical estimation may then be significantly more challenging.

Acknowledgements

The authors would like to thank Philipp Koehn for providing training scripts used in this work; and Steve Renals, Mirella Lapata and members of the Edinburgh SMT Group for valuable comments. This work was supported by an MRC Priority Area Studentship to the School of Informatics, University of Edinburgh.

References

Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. 2004. A probabilistic framework for semi-supervised clustering. In *Proc. of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*.

Julian Besag. 1986. The statistical analysis of dirty pictures. *Journal of the Royal Society Series B*, 48(2):259–302.

Peter Brown, Stephen Della Pietra, Vincent Della Pietra, and Robert Mercer. 1993. The mathematics of machine translation: Parameter estimation. *Computational Linguistics*, 19(2):263–311.

Philip A. Chou. 1991. Optimal partitioning for classification and regression trees. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(4).

M. Cmejrek, J. Curin, J. Havelka, J. Hajic, and V. Kubon. 2004. Prague Czech-English dependency treebank: Syntactically annotated resources for machine translation. In *4th International Conference on Language Resources and Evaluation, Lisbon, Portugal*

S. Geman and D. Geman. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 6:721–741.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proc. of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the HLT/NAACL 2003*.

Philipp Koehn. 2004. Pharaoh: a beam search decoder for phrase-based statistical machine translation models. In *Proceedings of the AMTA 2004*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT Summit 2005*.

S. Niessen and H. Ney. 2004. Statistical machine translation with scarce resources using morpho-syntactic information. *Computational Linguistics*, 30(2):181–204.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

F.-J. Och. 1998. An efficient method for determining bilingual word classes. In *Proc. of the European Chapter of the Association for Computational Linguistics 1998*.

Ye-Yi Wang, John Lafferty, and Alex Waibel. 1996. Word clustering with parallel spoken language corpora. In *Proc. of 4th International Conference on Spoken Language Processing, ICSLP 96, Philadelphia, PA*.

D.H. Wolpert. 1995. Determining whether two data sets are from the same distribution. In *15th international workshop on Maximum Entropy and Bayesian Methods*.

Mei Yang and Katrin Kirchhoff. 2006. Phrase-based backoff models for machine translation of highly inflected languages. In *Proc. of the the European Chapter of the Association for Computational Linguistics 2006*.

R. Zens and H. Ney. 2004. Improvements in phrase-based statistical machine translation. In *Proc. of the Human Language Technology Conference (HLT-NAACL 2004)*.