Mach Translat (2013) 27:85–114
DOI 10.1007/s10590-012-9132-2

N-gram posterior probability confidence measures for statistical machine translation: an empirical study

Adrià de Gispert · Graeme Blackwood · Gonzalo Iglesias · William Byrne

Received: 24 May 2011 / Accepted: 4 August 2012 / Published online: 28 August 2012

Abstract We report an empirical study of n-gram posterior probability confidence measures for statistical machine translation (SMT). We first describe an efficient and practical algorithm for rapidly computing n-gram posterior probabilities from large translation word lattices. These probabilities are shown to be a good predictor of whether or not the n-gram is found in human reference translations, motivating their use as a confidence measure for SMT. Comprehensive n-gram precision and word coverage measurements are presented for a variety of different language pairs, domains and conditions. We analyze the effect on reference precision of using single or multiple references, and compare the precision of posteriors computed from k-best lists to those computed over the full evidence space of the lattice. We also demonstrate improved confidence by combining multiple lattices in a multi-source translation framework.

Keywords Statistical machine translation · Minimum Bayes-risk decoding · Confidence measures · N-gram posterior probabilities

A. de Gispert · G. Iglesias · W. Byrne
Machine Intelligence Laboratory, Department of Engineering, Cambridge University, Cambridge, UK
e-mail: ad465@cam.ac.uk

G. Iglesias
e-mail: gi212@cam.ac.uk

W. Byrne
e-mail: wjb31@cam.ac.uk

G. Blackwood (✉)
IBM T.J. Watson Research, Yorktown Heights, NY 10598, USA
e-mail: blackwood@us.ibm.com

1 Introduction and motivation

We investigate the use of n-gram posterior probabilities as a confidence measure for statistical machine translation (SMT). Our empirical study demonstrates that the probability distribution based on the translation model and language model can be used to answer questions such as "with what probability does an n-gram occur in the reference translations?", "what percentage of words in a hypothesis can be expected to occur in the reference translations?", and "with what probability will a particular n-gram not be present in the references?"

We have previously shown that high posterior probability n-grams in the maximum likelihood translation hypothesis are more likely to be found in human reference translations (Blackwood et al. 2010b). This article builds upon those experiments to form a comprehensive study of n-gram posterior probability confidence measures for a range of different language pairs and evaluation frameworks. We first describe a practical and efficient algorithm that allows for rapid computation of n-gram posterior probabilities from SMT word lattices. We show that our algorithm enables efficient calculation of the n-gram posterior probabilities required for lattice minimum Bayes-risk (MBR) decoding and for confidence estimation.

Sentence-level and word-level confidence measures for SMT have been previously explored in Blatz et al. (2004), Ueffing and Ney (2005) and Ueffing and Ney (2007). Our approach is motivated by the greater flexibility and improved accuracy obtained by estimating confidence at the more fine-grained level of n-grams. We analyse in detail and for various conditions the power of posterior probabilities to predict whether certain parts of a given translation will or will not be found in human references. Our approach relies on statistics generated by the SMT-systems themselves, similar to Ueffing and Ney (2005) and Ueffing and Ney (2007). This is a somewhat different approach than the black-box approaches used in regression-based models (Specia et al. 2009a) which can operate without detailed information from the SMT systems. Of course, the statistics we produce could themselves be incorporated into regression, in the 'glass-box' modelling scenario (Specia et al. 2009b).

There are many potential applications of confidence measures estimated over n-grams. For example, in interactive MT (Casacuberta et al. 2009) and computer-aided translation (Barrachina et al. 2009), the process of translation iterates between confidence-based predictive translation and incorporation of user feedback until the desired level of quality is reached. Sentence-level confidence measures have been used to assign confidence to hypotheses in an interactive MT system (González-Rubio et al. 2010); our n-gram confidence measure could be used to more rapidly identify parts of the translation that are likely to require correction or refinement. Another potential use of n-gram confidence measures is error-driven source sentence paraphrasing (Buzek et al. 2010; Resnik et al. 2010), where the goal is to improve translation quality by identifying poorly translated fragments of the source sentence, and asking the user for a differently worded paraphrase that can be more easily and accurately translated. Confidence measures over n-grams can also be used to facilitate the targeted application of models intended to address particular deficiencies in SMT hypotheses, such as the monolingual coverage constraints of Blackwood et al. (2010b). In that approach, high confidence regions of the translation hypothesis are held fixed and more sophisticated models applied in re-decoding over the low confidence regions. Confidence measures could also be used to more effectively harvest user corrections in order to learn sources of error and improve translation models: it is not always easy for humans to correct an entire sentence if it contains many translation errors; it is better to ask users to first correct the substrings of the translation identified as being of low confidence. While detailed exploration of any of the above applications is beyond the scope of the analysis presented in this paper, we do approximate a post-editing scenario with the use of an automatic metric of translation quality, i.e. translation edit rate (TER) (Snover et al. 2006).

Confidence measures have also been extensively used for automatic speech recognition (ASR) (Rahim et al. 1997; Jiang and Huang 1998; Wessel et al. 2001; Jiang 2005). Note that the work of Wessel et al. (2001) is similar to the use of n-gram posteriors probabilities as an SMT confidence measure that we investigate in this article. Typical applications in ASR include utterance verification, correction of recognition results, detection or rejection of out-of-vocabulary words, and managing the flow of control in dialogue systems. One of the main differences between ASR and SMT confidence estimation is that the time-series nature of the ASR acoustic models tends to result in highly concentrated word-level confidence. Reordering in translation means that the confidence associated with translated words or phrases is often distributed over a range of possible target-language word positions, sometimes quite distant positions for language pairs requiring significant reordering.

The remainder of this article is structured as follows. We start by reviewing the linearized form of lattice MBR decoding and definition of n-gram posterior probabilities in Sect. 2. Then, in Sect. 3, we present an efficient algorithm for computing n-gram posterior probabilities from SMT word lattices. In Sect. 4, we define our confidence measure for single-lattice and multiple-lattice evidence spaces. Our Arabic→English, Chinese→English, French→English, and Spanish→English evaluation frameworks are described in Sect. 5. The efficiency of our n-gram posterior probability computation algorithm is evaluated in Sect. 6. This is followed by a detailed study of the predictive power of n-gram posterior probabilities as a confidence measure in Sect. 7. We include experiments examining the effect of translating from noisy or cleaned versions of the source-language input sentences, show the importance of the size of the evidence space in computing n-gram posterior probabilities, compare precisions computed with respect to single or multiple references, and evaluate the effect of multiple-lattice system combination on n-gram precision in a multi-source translation task. We conclude with a summary and discussion of related work in Sect. 8.

2 Lattice MBR decoding

MBR decoding can be applied to any MT system that defines a posterior distribution over translation hypotheses. MBR decoding for SMT (Kumar and Byrne 2004) has the general form in (1):

$$\hat{E} = \arg\min_{E' \in \mathcal{E}} \sum_{E \in \mathcal{E}} L(E, E') P(E|F), \tag{1}$$

where $\mathcal{E}$ is some space of translation hypotheses (e.g. a k-best list or lattice), $L(E, E')$ defines the loss between two hypotheses E and E', and $P(E|F)$ is the posterior probability of translating source sentence F as target sentence E. For a log-linear model of translation (Och and Ney 2002) the posterior has the form in (2):

$$P(E|F) = \frac{\exp(\alpha H(E, F))}{\sum_{E'} \exp(\alpha H(E', F))}, \tag{2}$$

where $H(E, F)$ is the score assigned by the model to sentence pair (E, F), typically the dot product of feature weights and feature values. The scaling factor α smoothes the posterior distribution, flattening when $\alpha < 1$ and sharpening when $\alpha > 1$.

The linearized form of the lattice MBR decoder (Tromble et al. 2008) replaces the loss function in Eq. (1) with a conditional expected gain based on an approximation to the BLEU score (Papineni et al. 2002). The conditional expected gain is computed as a weighted sum of local n-gram gain functions and a constant multiplied by the hypothesis length to give (3):

$$\hat{E} = \arg\max_{E' \in \mathcal{E}} \left\{ \theta_0 |E'| + \sum_{n=1}^{4} \sum_{u \in \mathcal{N}_n} \theta_n \#_u(E') p(u|\mathcal{E}) \right\}, \tag{3}$$

where $\mathcal{E}$ is a lattice of translation hypotheses, $\mathcal{N}_n$ is the set of n-grams (of order n) in the lattice, $\#_u(E')$ is the number of times the n-gram u occurs in hypothesis E', and the parameters θ are constants estimated on held-out data as described in Sect. 5.1 of Tromble et al. (2008). The quantity $p(u|\mathcal{E})$ is the path posterior probability of the n-gram u. This posterior is defined as in (4):

$$p(u|\mathcal{E}) = \sum_{E \in \mathcal{E}} \delta_u(E) P(E|F) = \sum_{E \in \mathcal{E}_u} P(E|F), \tag{4}$$

where $\delta_u(E)$ has the value 1 if u occurs in E and 0 otherwise, and $\mathcal{E}_u = \{E \in \mathcal{E} : \#_u(E) > 0\}$ is the subset of paths containing the n-gram u at least once. We note that the posterior probability of Eq. (4) differs from the expected count $c(u|\mathcal{E}) = \sum_{E \in \mathcal{E}} \#_u(E) P(E|F)$ since probability must be accumulated once per path rather than in proportion to the number of times the n-gram occurs on each path (Blackwood et al. 2010a). Throughout this paper we will use n-gram posteriors computed using path posterior probabilities.

Equation (3) approximates the general form of the MBR decoder in Eq. (1) by replacing the sum over all paths in the lattice by a sum over lattice n-grams. Although a lattice may have many n-grams, it is possible to extract and enumerate these n-grams exactly, whereas it is usually impossible to enumerate all paths. Therefore, while the linearisation of the gain function in the decision rule is an approximation, it has the advantage that Eq. (3) can be computed exactly even for very large lattices.

Section 4 will show that the n-gram posterior probabilities used to compute the conditional expected gain are a good predictor of whether or not an n-gram is to be found in the reference translations. This motivates their use as an n-gram confidence measure for SMT. Before investigating their use in confidence estimation, we first describe a practical and efficient algorithm for exact computation of n-gram posterior probabilities from a lattice.

3 Efficient posterior probability computation from translation lattices

We now describe procedures for the efficient computation of n-gram posteriors from translation lattices. An MT word lattice (Ueffing et al. 2002) has the form of a directed acyclic graph (Cormen et al. 2001); we encode the word sequences and scores of translation hypotheses in the lattice as a weighted finite-state acceptor (WFSA) (Kumar and Byrne 2003). The experiments we report in this paper are based on HiFST (Iglesias et al. 2009b), which directly generates dense representations of word-level hypotheses as WFSAs. This decoder has been described previously (de Gispert et al. 2010; Iglesias et al. 2011). It is particularly efficient in its representation of translation hypotheses and thus has natural advantages for generation of rich hypothesis spaces over which posterior probabilities can be computed.

Algorithms based on WFSA operations (Mohri 1997) have been previously developed to compute n-gram posterior probabilities from a lattice (Tromble et al. 2008). This approach can be slow for large lattices with many n-grams since it requires a separate intersection and summation over matching paths for each n-gram in the lattice. If we use weighted finite-state transducers (WFSTs) instead of acceptors, then the posterior probabilities of all n-grams of the same order can be computed in a single composition (Blackwood et al. 2010a). In this section, we present an efficient algorithm based on the forward procedure that allows for fast and exact computation of n-gram posterior probabilities. This algorithm is a lattice specialization of the hypergraph vector-indexed algorithm of DeNero et al. (2010).

Formally, a WFST $T = (\Sigma, \Delta, Q, I, F, E, \lambda, \rho)$ over a semiring $(\mathbb{K}, \oplus, \otimes, \bar{0}, \bar{1})$ is defined by an input alphabet Σ, an output alphabet Δ, a set of states Q, a set of initial states $I \subseteq Q$, a set of final states $F \subseteq Q$, a set of weighted transitions E, an initial state weight assignment $\lambda : I \rightarrow \mathbb{K}$, and a final state weight assignment $\rho : F \rightarrow \mathbb{K}$ (Mohri et al. 2008). The weighted transitions of T form the set $E \subseteq Q \times \Sigma \times \Delta \times \mathbb{K} \times Q$, where each transition includes a source state from Q, an input symbol from Σ, an output symbol from Δ, a cost from the weight set $\mathbb{K}$, and a target state from Q. For each state $q \in Q$, let $E[q]$ denote the set of edges leaving state q. For each transition $e \in E[q]$, let $p[e]$ denote its source state, $n[e]$ its target state, $i[e]$ its input label, $o[e]$ its output label, and $w[e]$ its weight. A WFSA can be derived from a WFST by projecting on the input or output labels; this operation is denoted $\Pi_1(T)$ for input projection and $\Pi_2(T)$ for output projection.

The overall approach can be summarized as follows. The core of the approach is the modified forward procedure mentioned above and which is detailed in Fig. 1. The typical forward procedure calculates forward probabilities $\alpha(q)$: this is the marginal probability of the partial paths which lead from the start state to state q. By contrast, the modified forward procedure of Fig. 1 calculates quantities $\alpha(q, u)$: these are the marginal probabilities of the paths which lead to state q and that pass through at least one arc with the input symbol u. Somewhat obviously, these differ from the usual

```
Compute-Ngram-Posteriors
for each state $q \in Q$   ▷ In topologically sorted order
    do for each edge $e \in E[q]$
        do $\alpha(n[e]) \leftarrow \alpha(n[e]) + (\alpha(q) \times w[e])$
            if $i[e] \notin \mathcal{N}_{n[e]}$
                then $\mathcal{N}_{n[e]} \leftarrow \mathcal{N}_{n[e]} \cup \{i[e]\}$
            $\alpha(n[e], i[e]) \leftarrow \alpha(n[e], i[e]) + (\alpha(q) \times w[e])$
            for each $n$-gram $u \in \mathcal{N}_q$ where $u \neq i[e]$
                do if $u \notin \mathcal{N}_{n[e]}$
                        then $\mathcal{N}_{n[e]} \leftarrow \mathcal{N}_{n[e]} \cup \{u\}$
                    $\alpha(n[e], u) \leftarrow \alpha(n[e], u) + (\alpha(q, u) \times w[e])$
    if $q \in F$
        then for each $n$-gram $u \in \mathcal{N}_q$
                do $p(u|\mathcal{E}) \leftarrow p(u|\mathcal{E}) + (\alpha(q, u) \times \rho[q])$
    $\mathcal{N}_q \leftarrow \emptyset$   ▷ Clean up state $q$
```

Fig. 1 Algorithm for efficient computation of n-gram posteriors based on path-posterior probabilities. The input to the procedure is an order-n mapped lattice $\mathcal{E}_n$ that is assumed to be topologically sorted and normalized such that each path is weighted with its posterior probability $P(E|F)$. The output of the procedure is the posterior probability $p(u|\mathcal{E})$ of each n-gram in $\mathcal{E}_n$. $\mathcal{N}_q$ denotes the set of all symbols observed on partial paths to state q

words	bigram
$w_1 w_1$	u_1
$w_2 w_2$	u_2
$w_1 w_2$	u_3
$w_2 w_1$	u_4

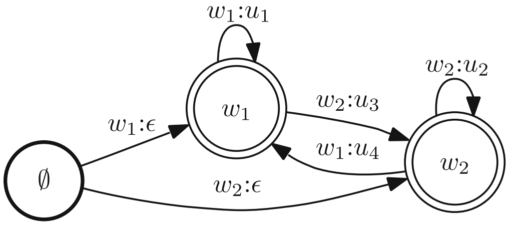

Fig. 2 Mapping transducer Φ_2 for all possible bigrams $\Sigma_2 = \{u_1, u_2, u_3, u_4\}$ formed from unigram alphabet $\Sigma_1 = \{w_1, w_2\}$. *States* and *arcs* need only be added for bigrams $u \in \mathcal{N}_2$

forward probabilities in that $\alpha(q, u) \leq \alpha(q)$ and $\sum_u \alpha(q, u) \geq \alpha(q)$. This algorithm should properly be thought of as a modified form of marginalization, rather than a counting procedure, and it yields statistics for the calculation of Eq. (4) for the case when u are unigrams. We could extend the modified forward procedure to marginalize probabilities over paths which contain n-grams. However, it is easier first to transduce word lattices to n-gram lattices and then use the modified forward procedure simply to count individual n-gram tokens, as we describe next.

Let Φ_n denote the mapping transducer for n-grams of order n. This transducer maps word sequences to sequences of n-grams so that the order-n mapped lattice $\mathcal{E}_n$ is obtained by composing the word lattice $\mathcal{E}$ with Φ_n, projecting on the output, removing ϵ-transitions, determinizing and minimizing, as in (5):

$$\mathcal{E}_n = \min(\det(\text{rmeps}(\Pi_2(\mathcal{E} \circ \Phi_n)))) \tag{5}$$

The resulting acceptor $\mathcal{E}_n$ is a compact lattice of n-gram sequences of order-n consistent with the hypotheses and scores of the original lattice $\mathcal{E}$. Figure 2 shows an example mapping transducer Φ_2 for any sequence of bigrams in $\{u_1, u_2, u_3, u_4\}^*$ that can be constructed from the unigram alphabet $\Sigma_1 = \{w_1, w_2\}$. The input labels correspond to matched word sequences; the output labels define the corresponding transduced bigram sequences. As a simple example, applying Φ_2 to the word sequence $w_1\, w_2\, w_1$ yields the bigram sequence $u_3\, u_4$.

The algorithm shown in Fig. 1 computes in a single forward pass the posterior probabilities $p(u|\mathcal{E})$ of all n-grams u in an order-n mapped lattice $\mathcal{E}_n$. We assume that the lattice has been topologically sorted and normalized [by pushing weights in the log semiring (Mohri et al. 2008)] such that a path labeled with the words of hypothesis E has weight $P(E|F)$ in accordance with Eq. (2); forward probabilities α are initialized to 0. The order-n mapping transduction and forward algorithm are repeated for each order $n = 1 \ldots 4$. After computing the posterior probabilities $p(u|\mathcal{E})$ of each n-gram $u \in \mathcal{N}$, we perform lattice MBR decoding over the full lattice using the implementation described in Blackwood et al. (2010a). We report the efficiency of our new symbol-specific forward algorithm in the lattice MBR decoding experiments of Sect. 6.

4 N-gram posterior probability confidence measures

We are interested in the predictive power of n-gram posterior probabilities. We would like to analyze the relation between posterior probability and translation quality by computing: (a) the precision of high posterior n-grams with respect to the human reference translations available for each source sentence; (b) the translation hypothesis coverage of high posterior n-grams; (c) the converse precision of low posterior n-grams with respect to the human references; and (d) the precision of high posterior n-grams in a system combination scenario. This section describes how we compute these.

4.1 Posterior probability reference precisions

For each sentence, let $\mathcal{N}_n$ denote the set of n-grams of order n for which we would like to compute the reference precision. This could be the n-grams of the ML translation 1-best hypothesis $\hat{E}$ or the n-grams in a subset of translation hypotheses such as a k-best list or lattice; in this paper we focus on n-grams present in the ML translation 1-best hypothesis. Let $\mathcal{R}_n$ denote the set of n-grams of order n in the union of the references. For confidence threshold β, let $\mathcal{N}_{n,\beta} = \{u \in \mathcal{N}_n : p(u|\mathcal{E}) \geq \beta\}$ denote the set of n-grams in $\mathcal{N}_n$ with posterior probability greater than or equal to β, where $p(u|\mathcal{E})$ is computed according to Eq. (4) from the translation lattice $\mathcal{E}$. The precision at order n for threshold β is the proportion of n-grams in $\mathcal{N}_{n,\beta}$ also present in the references, as in (6):

$$\mathcal{P}_{n,\beta} = \frac{|\mathcal{R}_n \cap \mathcal{N}_{n,\beta}|}{|\mathcal{N}_{n,\beta}|} \tag{6}$$

4.2 Posterior probability hypothesis coverage

To complement the precision scores above, we investigate how many words in the top hypothesis are covered by $\mathcal{N}_{n,\beta}$ at each confidence threshold β. Let I be the length of the ML translation 1-best hypothesis, and let $\mathcal{W}_{n,\beta}$ denote the set of words in the hypothesis that belong to n-grams of order n with posterior probability greater than or equal to β. Then the coverage at order n for threshold β is the proportion of hypothesised words covered by n-grams in $\mathcal{N}_{n,\beta}$, as in (7):

$$\mathcal{C}_{n,\beta} = \frac{100 * |\mathcal{W}_{n,\beta}|}{I - n + 1} \tag{7}$$

We note that this notion of coverage could also be defined with respect to the length of the available reference(s). Here we choose to calculate a well-defined percentage over hypothesised words and simply report the BLEU score brevity penalty for each of the sets we will analyse. This is because the brevity penalty reflects any divergence in length between the hypothesis and the closest (in length) of the available references.

We also note that the above coverage is defined on a single hypothesis, but it could easily be extended to average scores over a k-best list or even a lattice. However, we will use this simpler formulation in our experiments section, as we believe it is more intuitive and makes discussion of the predictive power of posteriors easier.

4.3 Posterior probability converse reference precisions

Additionally, for confidence threshold γ, let $\mathcal{N}_{n,\gamma} = \{u \in \mathcal{N}_n : p(u|\mathcal{E}) \leq \gamma\}$ denote the set of n-grams in $\mathcal{N}_n$ with posterior probability *lower* than or equal to γ. The *converse* precision at order n for threshold γ is the proportion of n-grams in $\mathcal{N}_{n,\gamma}$ that are *not* present in the references, as in (8):

$$\mathcal{Q}_{n,\gamma} = \frac{|\mathcal{N}_{n,\gamma} \setminus \mathcal{R}_n|}{|\mathcal{N}_{n,\gamma}|} \tag{8}$$

This quantity tests the ability of the posteriors to indicate when portions of the hypothesised translation cannot be trusted; ideally, low posterior values should be as informative and predictive as high posterior values.

4.4 System combination reference precisions

Finally, we also consider the effect on reference precision of computing n-gram posterior probabilities from a combination of multiple translation lattices in the context of multi-input and multi-source translation (Blackwood 2010). Treating each lattice as a WFSA, we denote the evidence space formed by taking the union of M individual translation lattices $\mathcal{E}^{(1)}, \ldots, \mathcal{E}^{(M)}$ as $\mathcal{E}$. The posterior probability of n-gram u according to the evidence space of lattice $\mathcal{E}^{(i)}$ is computed as in (9):

$$p_i(u|\mathcal{E}^{(i)}) = \sum_{E \in \mathcal{E}_u^{(i)}} P(E|F), \tag{9}$$

where $\mathcal{E}_u^{(i)} = \{E \in \mathcal{E}^{(i)} : \#_u(E) > 0\}$ denotes the set of all paths in lattice $\mathcal{E}^{(i)}$ with one or more occurrences of the n-gram u. We compute the n-gram confidence $p(u|\mathcal{E})$ as a weighted combination of the individual lattice n-gram posterior probabilities of Eq. (9) using one of the following two methods:

Weighted sum The confidence of n-gram u computed according to a weighted sum of the posterior probabilities in each of the M lattices is (10):

$$p(u|\mathcal{E}) = \sum_{i=1}^{M} \lambda_i \; p_i(u|\mathcal{E}^{(i)}), \quad 0 \le \lambda_i \le 1, \quad \sum_{i=1}^{M} \lambda_i = 1. \tag{10}$$

Weighted product The confidence of n-gram u computed according to a weighted product of the posterior probabilities in each of the M lattices is (11):

$$p(u|\mathcal{E}) \propto \prod_{i=1}^{M} p_i(u|\mathcal{E}^{(i)})^{\lambda_i}, \quad 0 \le \lambda_i \le 1, \quad \sum_{i=1}^{M} \lambda_i = 1 \tag{11}$$

where the posteriors are normalized to sum to 1.

The values of the λ_i^M in Eqs. (10) and (11) should reflect the quality of the various systems. We set the values by performing a grid-based search over the parameter values: at each grid point we perform LMBR-based system combination, and we choose the grid value based on optimal BLEU score on a development set (Blackwood 2010). This is standard practice for tuning system combination (Rosti et al. 2007; Sim et al. 2007), and we note that it can be done very quickly based on lattices already generated by the contributing systems.

5 System development

We investigate lattice MBR decoding efficiency and the predictive power of n-gram posterior probabilities as a confidence measure for a range of language pairs and experimental conditions. This section describes the baseline translation systems and MT pipeline used for the experiments reported in Sects. 6 and 7. Our n-gram posterior probability computation algorithm and lattice MBR decoder are implemented using OpenFST (Allauzen et al. 2007).

5.1 Arabic→English and Chinese→English translation

Arabic→English experiments are carried out for the NIST MT09 translation task.[1] The development set mt0205tune is formed from the odd-numbered sentences of the MT02–MT05 evaluation sets; the even-numbered sentences form the test set mt0205test. For Chinese→English translation, the testsets are from the GALE P3

[1] http://www.itl.nist.gov/iad/mig/tests/mt/2008/.

Table 1 Arabic→English and Chinese→English evaluation framework testset statistics and genres

Language	Testset	Genre	# Sentences	Avg Src length
AR→EN	mt0205tune	news	2,075	31.2
	mt0205test	news	2,040	31.3
ZH→EN	tune.nw	news	1,755	30.9
	tune.web	web	2,495	27.6

evaluation and include both newswire and web data. The testset statistics and genres are summarized in Table 1.

Word alignments for both language pairs are generated using MTTK (Deng and Byrne 2008). For Arabic→English translation, the alignments are generated over approximately 150M words of parallel text specified for the constrained NIST MT09 Arabic→English track. Prior to generating the alignments, the Arabic side of the parallel text is pre-processed with the MADA morphological toolkit (Habash and Rambow 2005). The word alignments for Chinese→English translation are trained from around 250M words of parallel text distributed for the GALE P3 evaluation.

Hierarchical rules are extracted from the aligned text using the constraints described in Chiang (2007) with the count and pattern filters of Iglesias et al. (2009a). First-pass translation decoding with HiFST (Iglesias et al. 2009b) generates word lattices encoding large numbers of alternative hypotheses. Minimum error rate training (Och 2003) under the BLEU score (Papineni et al. 2002) optimizes the following features with respect to the development set: target language model, source→target and target→source rule translation probabilities, word and rule penalties, number of usages of the glue rule, source→target and target→source lexical translation probabilities, and three count-based features that track the frequency of rules in the parallel data (Bender et al. 2007). The English language model used during decoding is a modified Kneser-Ney (Kneser and Ney 1995) smoothed 4-gram estimated over the English side of the parallel text and a 465M word subset of the English GigaWord Third Edition (Graff et al. 2007).

First-pass lattices are rescored with 5-gram sentence-specific zero-cutoff stupid-backoff language models (Brants et al. 2007) estimated over more than six billion words of English language training text. The scaling parameter α and per-word factor θ_0 in Eq. (3) are optimized with respect to the appropriate development set: mt0205tune for Arabic→English translation experiments, and tune.nw or tune.text.web for Chinese→English translation experiments.

5.2 French→English translation

French→English reference precisions are evaluated in the context of the Cambridge University Engineering Department (CUED) hierarchical phrase-based SMT pipeline, as submitted to the ACL Fifth Workshop on Statistical Machine Translation 2010 Shared Task.[2] Our WMT 2010 translation pipeline is similar to that used for

[2] http://www.statmt.org/wmt10.

Table 2 Parallel training data statistics for WMT 2010 French→English experiments

Language	# Sentences	# Tokens	# Types
English	30.2M	962.4M	2.4M
French	30.2M	815.3M	2.7M

Arabic→English and Chinese→English translation described above; full details are available in our CUED WMT 2010 shared task system description paper (Pino et al. 2010).

For the experiments reported below, first-pass translation lattices are generated by the HiFST decoder using a Shallow-1 grammar (de Gispert et al. 2010) estimated from all of the available parallel data (summarized in Table 2). There is no pruning during first-pass decoding. Separate 4-gram Kneser-Ney smoothed language models (Kneser and Ney 1995) are estimated for each of the monolingual corpora. A single interpolated model is built by optimizing the weights of each component in order to minimize perplexity on the development set. Prior to computing n-gram posterior probabilities using Eq. (4), first-pass lattices are rescored with zero-cutoff stupid-backoff language models (Brants et al. 2007) estimated over the English side of the parallel text and additional monolingual data, a total of around 6.2 billion tokens. The shared task newstest2008, newstest2009, and newstest2010 sets are used to evaluate the precision of n-grams in the ML first-pass 1-best translations.

5.3 Spanish→English and English→Spanish translation

The goal of the FP7 Feedback Analysis for User Adaptive Statistical Translation (FAUST) project is to develop MT systems that respond rapidly and intelligently to user feedback. The main objectives of the project are (i) to identify and immediately incorporate useful feedback in the development and evaluation cycle of systems deployed for online translation, and (ii) to improve user satisfaction with online MT by integrating natural language generation to improve translation fluency.

FAUST includes translation from both noisy and cleaned versions of source-language sentences. Separate development and test sets are available for multiple language pairs. In this article, we consider the Spanish→English and English→Spanish translation tasks. The ‘dev’ and ‘test’ testsets were created from a collection of online translation requests at Reverso.[3] Cleaned versions of the original source sentences were prepared by two independent translators,[4] This mainly involved minor changes to the original requests, such as spelling and grammar correction. The ‘clean’ version is meant to be unambiguously translatable; this assessment was left to the judgement of each translator. Note that if the original source request was unambiguously

[3] http://www.reverso.net.

[4] In response to the view that extending our analysis to the material generated within the FAUST project might be redundant, we would like to stress that this is real-world data, taken from web server logs from Europe’s most heavily trafficked online provider of free translation services. This data is what users are actually submitting for translation. We believe that validating the results of our research with well-studied static corpora on this type of material is of great importance.

Table 3 Parallel training data statistics for FAUST Spanish↔English experiments

Language	# Sentences	# Tokens	# Types
English	1.7M	49.4M	167.2K
Spanish	1.7M	47.0M	122.7K

translatable, it was taken as the ‘clean’ version. Each translator then generated a reference translation from the cleaned source-language request. The FAUST data sets are available on the project website.[5]

Development and test sets containing 1,000 sentences were created where each sentence includes the original noisy source, two versions of the cleaned source, and two human reference translations. We denote the original source by ‘os’. The cleaned source and reference translation produced by translator *n* are denoted ‘cs*n*’ and ‘rt*n*’, respectively ($n = 0, 1$). Two English→Spanish translation examples from the development set are shown below:

os the things you say behind my back say it to my face and let that be the end of it.
cs0 the things you say behind my back, say them to my face and let that be the end of it.
cs1 the things you say behind my back say to my face and let that be the end of it.
rt0 las cosas que me dices por detrás, dímelas a la cara y acabemos con esto.
rt1 las cosas que dices a mi espalda dímelas a la cara y acabemos con esto de una vez.

os present this flyer on your first lesson and receive your first hour free.
cs0 present this flyer at your first lesson and get your first hour free.
cs1 present this flyer on your first lesson and receive your first hour free.
rt0 presenta este flyer en tu primera clase y tendrás la primera hora gratis.
rt1 presenta este folleto en tu primera clase y recibirás tu primera hora gratis.

Lattice MBR decoding performance and *n*-gram posterior probability reference precisions are evaluated for Spanish↔English translation. First-pass lattices and 1-best translations are generated with the same hierarchical phrase-based decoder as used in the CUED submission to the WMT 2010 Spanish↔English shared translation task (Pino et al. 2010). The parallel training data used to train the baseline system is summarized in Table 3. The first-pass target LM for both Spanish→English and English→Spanish translation is estimated over the relevant side of the parallel data together with the additional monolingual ‘News’ training data distributed for WMT 2010. The feature weights of the decoder are tuned to optimize BLEU on the cleaned FAUST dev set using MERT (Och 2003).

In the FAUST experiments, lattice MBR decoding is applied directly to the first-pass HiFST translation lattices, i.e. there is no 5-gram language model rescoring prior to lattice MBR. In this article, BLEU-2 scores and reference precisions are computed

[5] http://www.faust-fp7.eu/faust.

Table 4 Average time (s/sentence) to compute n-gram path posterior probabilities using the sequential method, path counting transducers, and symbol-specific forward algorithm

	Arabic→English		Chinese→English	
	mt0205tune	mt0205test	tune.nw	tune.web
Sequential	1.52	1.62	4.43	4.73
Transducers	0.84	0.88	1.68	1.69
Symbol-specific	0.13	0.14	0.41	0.40

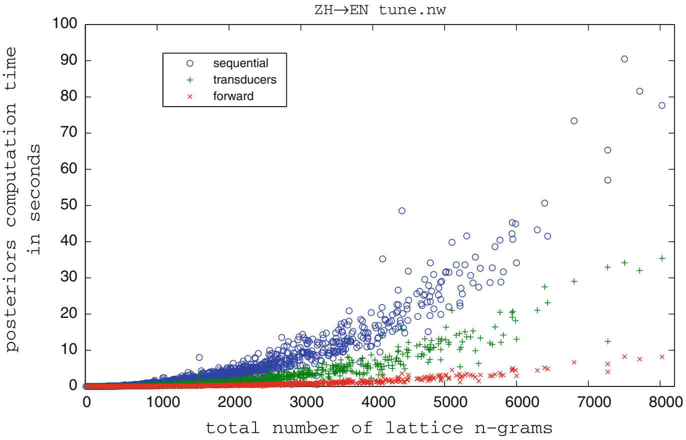

Fig. 3 Posterior probability computation time (s) versus # of lattice n-grams using the sequential method, path counting transducers, and symbol-specific forward algorithm for each sentence of the Chinese→English tune.nw testset

with respect to the union of the rt0 and rt1 references using case-insensitive n-gram matching (denoted ‘rt0,1’).

6 MBR decoding efficiency

We compare the efficiency of lattice MBR decoding using the sequential method (Tromble et al. 2008), path counting transducers (Blackwood et al. 2010a), and the symbol-specific forward algorithm described in Sect. 3.

The time required for lattice MBR decoding is dominated by the time required to compute the n-gram posterior probabilities. Table 4 shows the average time in seconds per sentence required to compute these statistics for the NIST MT09 Arabic→English and GALE P3 Chinese→English testsets. These results show that the symbol-specific forward algorithm is several times faster than the implementation using path counting transducers, and more than an order of magnitude faster than the original sequential implementation. The average time required to compute n-gram posteriors from a lattice is less than half a second per sentence, fast enough for inclusion in a real-time interactive or computer-aided translation system.

Figure 3 plots posterior probability computation time as a function of the number of lattice n-grams for the Chinese→English newswire tune.nw testset. This compares

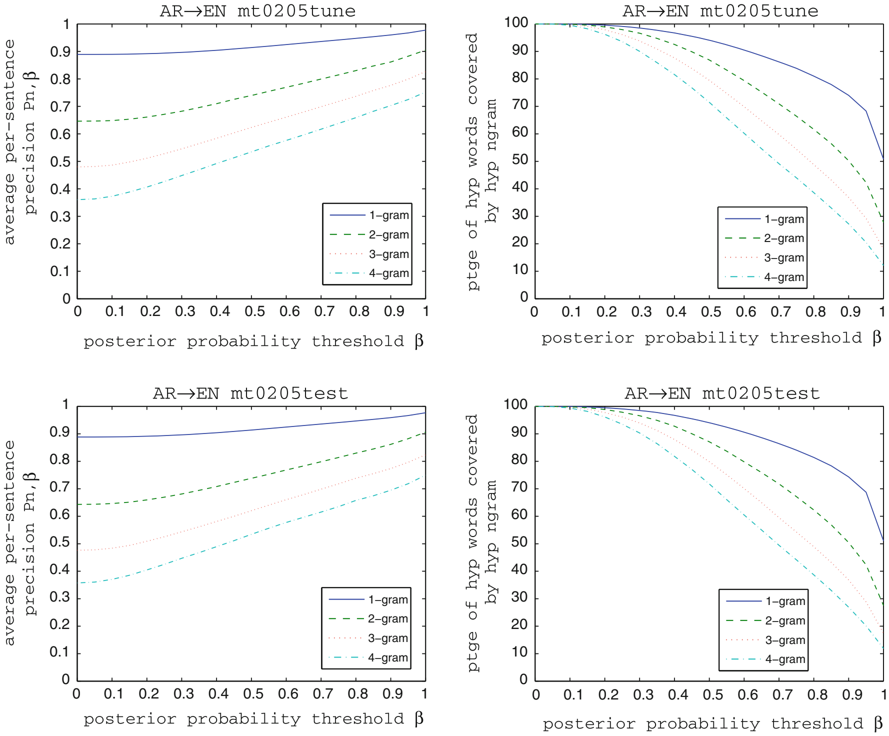

Fig. 4 Average per-sentence n-gram precisions and hypothesis coverage for Arabic→English 1-best ML translations of tune (*top*) and test (*bottom*) sets at a range of posterior probability thresholds $0 \leq \beta \leq 1$

the three methods on a per-sentence basis. The symbol-specific forward algorithm is clearly much more efficient than the other methods, particularly for longer sentences with many n-grams. The symbol-specific forward algorithm is used for all of the lattice MBR and confidence measure experiments described in the following sections.

7 Reference precision and confidence measure experiments

This section investigates the predictive power of n-gram posterior probabilities by computing reference precision and hypothesis coverage of high confidence n-grams, as well as converse reference precision of low confidence n-grams, as described in Sect. 4.

7.1 Arabic→English and Chinese→English

Arabic to English The left-most plots in Fig. 4 show average per-sentence n-gram precisions $\mathcal{P}_{n,\beta}$ by n-gram order for Arabic→English mt0205tune and mt0205test ML 1-best translations over a range of posterior probability thresholds $0 \leq \beta \leq 1$.

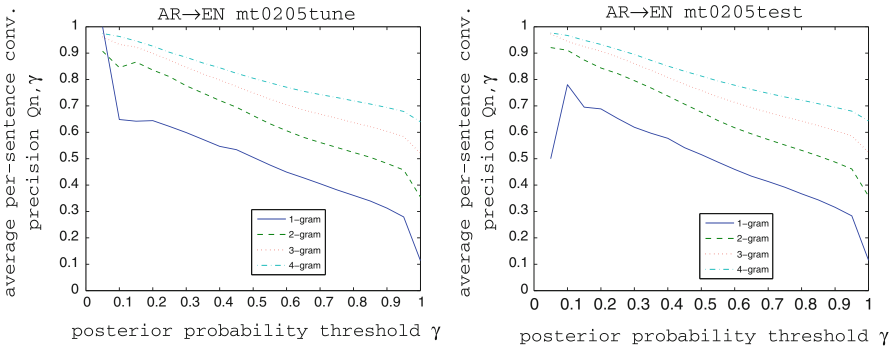

Fig. 5 Average per-sentence converse n-gram precisions for Arabic→English 1-best ML translations at a range of posterior probability thresholds $0 \leq \gamma \leq 1$

The BLEU scores for these two sets are 54.2 and 53.8, respectively, and the brevity penalty is just over 0.994, i.e. the hypotheses are 99.4 % as long as the closest reference. Sentence start and end tokens are ignored when computing unigram precisions. The plots show that precisions at all orders improve considerably as the threshold β increases, confirming that these intrinsic measures of translation confidence have strong predictive power. Note that the upper plots show at $\beta = 0$ the n-gram precisions used to compute the BLEU score of the ML 1-best translations.

The right-most plots show the percentage of words in the translation that are covered by n-grams with $p(u|\mathcal{E}) \geq \beta$, for the same range of β values. For high β, the percentage of words covered is smaller; this is as expected as there are fewer n-grams with high posteriors. However, even at a high threshold of $\beta = 0.9$, we are still on average covering ∼30 % of each translated sentence with high-confidence 4-grams, ∼40 % with 3-grams, ∼50 % with 2-grams and ∼75 % with 1-grams. Coupled with the precisions shown above, this is very useful information. For example, posteriors allow us to predict that 75 % of the words in a translation have a 0.96 probability of being present in the human references, irrespective of their order in the sentence. This coverage increases up to 95 % of the words for a 0.90 probability. For 2-grams, we can predict that 50 % of the words in the sentence will occur in the references with a probability of 0.86 (or 80 % for 0.77), and that they do so in sequences of two words. For 3-grams, we can predict that 43 % of the words in the sentence will be in the references with a probability of 0.76, and that they will occur in sequences of three words, and so on.

Complementarily to the above, the plots in Fig. 5 show average per-sentence converse n-gram precisions $\mathcal{Q}_{n,\beta}$ by n-gram order for Arabic→English mt0205tune and mt0205test ML 1-best translations over a range of posterior probability thresholds $0 \leq \gamma \leq 1$. The plots show that n-grams with a low posterior probability are most likely *not* to be found in the references. For example, 2-grams with a posterior lower than 0.2 (or 0.3) are not in the references in 84 % (or 78 %) of the cases. Interestingly, converse n-gram precisions are greater for high-order n-grams, so 4-grams with a posterior lower than 0.3 are not in the references in 88 % of the cases.

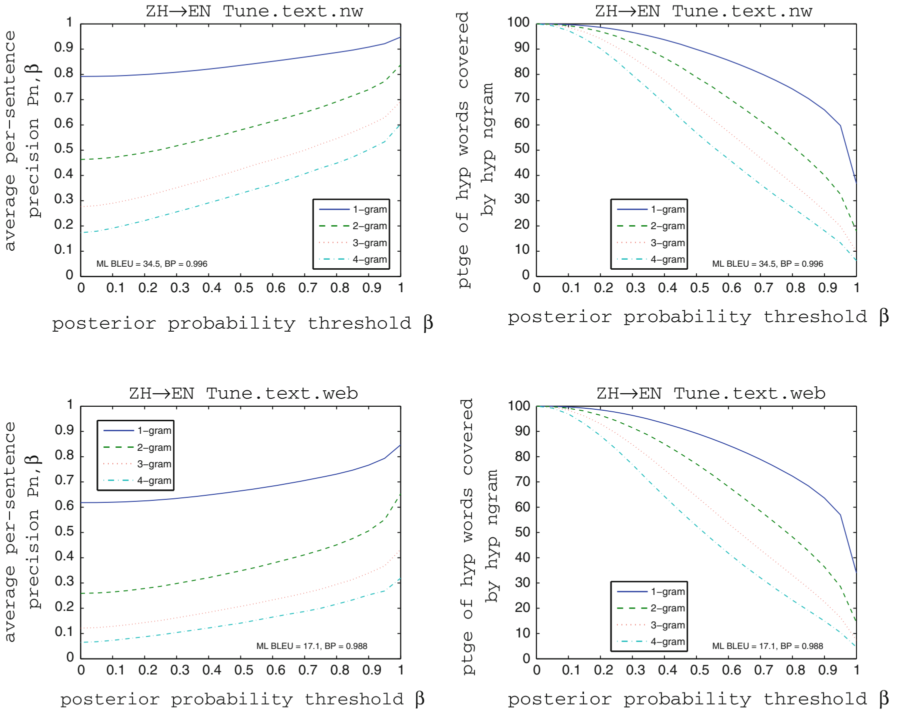

Fig. 6 Average per-sentence n-gram precisions and hypothesis coverage for Chinese→English newswire (*top*) and web data (*bottom*) 1-best ML translations at a range of posterior probability thresholds $0 \leq \beta \leq 1$

Note that both mt0205tune and mt0205test display similar trends in precision and coverage. We have observed this similarity among tune and test sets across all language pairs, domains and data conditions. Therefore, in the subsequent sections we report results for only one development set per condition.

Chinese to English Precision and coverage plots for GALE Chinese→English newswire (top) and web (bottom) data translations are shown in Fig. 6. The BLEU score is 34.5 (17.1 for web) and the brevity penalty is 0.996 (0.988 for web). Precisions at all orders are considerably lower than those for Arabic→English translation. This is unsurprising given the much lower BLEU score of the ML translations. The web data translations in particular have very low 3-gram and 4-gram precisions. In contrast, the converse precisions of low-confidence n-grams are significantly higher, indicating the quality prediction value of posteriors. This is shown in Fig. 7, where we see that most 4-grams with posterior lower than 0.4 will not occur in the references, especially in the web domain.

7.1.1 Relating n-gram posteriors and post-edition via TER

We further analyse the use of n-gram posteriors as confidence measures in the following automatic post-edition scenario. We first compute the TER score of the ML

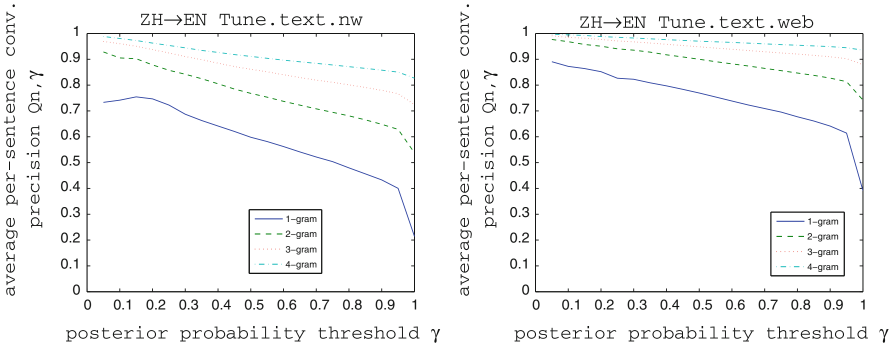

Fig. 7 Average per-sentence converse n-gram precisions for Chinese→English 1-best ML translations at a range of posterior probability thresholds $0 \leq \gamma \leq 1$

1-best translation. This is obtained by considering the number of deletions, insertions, substitutions and shifts required to edit the translation so that it matches entirely any one of the available references, as expressed by Eq. (12).

$$\text{TER} = \frac{\#\text{Dels} + \#\text{Ins} + \#\text{Subs} + \#\text{Shifts}}{\#\text{Words in Ref}} \tag{12}$$

Rather than the actual TER score, here we are more interested in the sequence of edits carried out automatically in order to compute it. These edits can be interpreted as a pessimistic approximation to what a human post-editor would do in attempting to amend the translation. It is pessimistic since there are typically many other correct translations than those available as golden references, and so one would expect a human post-editor to require fewer edits.

As before, for each confidence threshold β, $\mathcal{N}_{n,\beta}$ is the set of n-grams in the ML 1-best translation with posterior probability greater than or equal to β. Here we divide $\mathcal{N}_{n,\beta}$ into three subsets: (a) n-grams that are found in the TER reference, (b) n-grams that disappear due to shifts in editing the translation, and (c) n-grams that disappear due to deletions, substitutions and insertions. We report these in percentages over $\mathcal{N}_{n,\beta}$ for Arabic→English in Fig. 8 for 1-gram, 2-gram, 3-gram and 4-gram. As shown, the proportion of hypothesised n-grams that require edits decreases as the confidence threshold goes up.

To illustrate this, consider the two examples shown in Fig. 9, where a translation hypothesis (H) is shown, together with the TER edits required to reach a reference translation (R). H$'$ is the reordered hypothesis after editing shifts. If a hypothesis n-gram appears in (R), then it is correct and belongs to our subset (a). Otherwise, if it does not appear in (R), then it either falls into (b) or (c); if the n-gram does not appear in (H$'$), i.e. it was lost due to shifting, it belongs to (b); and if the n-gram appears in (H$'$) but not in (R), i.e. it was edited in place, it belongs to (c). In the left-most example, the only 3-gram with a posterior greater than 0.8 is 'The international federation', which is found in the TER reference. The 2-grams with posterior greater than 0.8 are 'The international', 'international federation', and 'suspended as'; only two of those appear in the reference, as the latter disappears due to substitutions. In the right-most

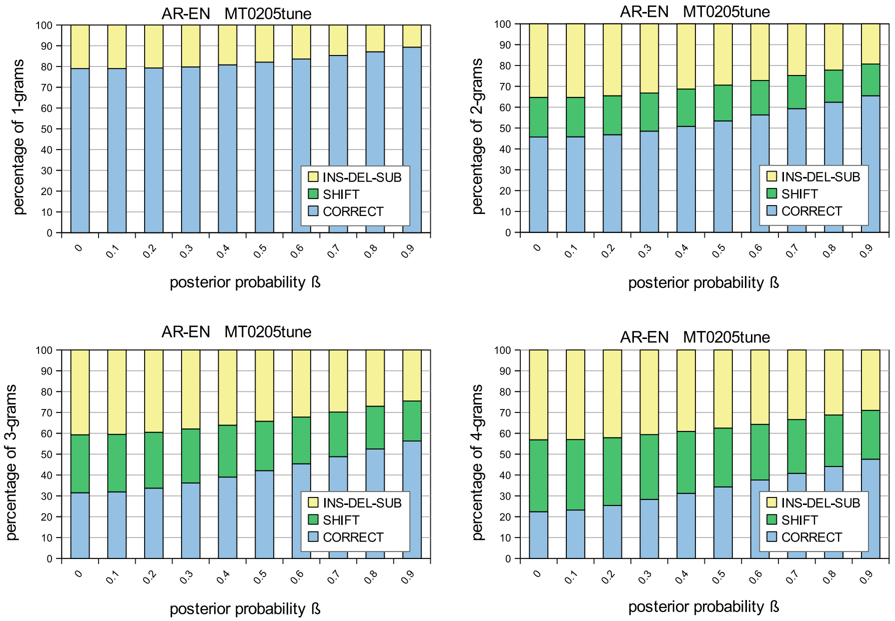

Fig. 8 Average percentage of Arabic→English 1-best ML translation 1-grams, 2-grams, 3-grams and 4-grams that are deemed correct by TER, that disappear due to both shifts and other edits at a range of posterior probability thresholds $0 \leq \beta \leq 1$

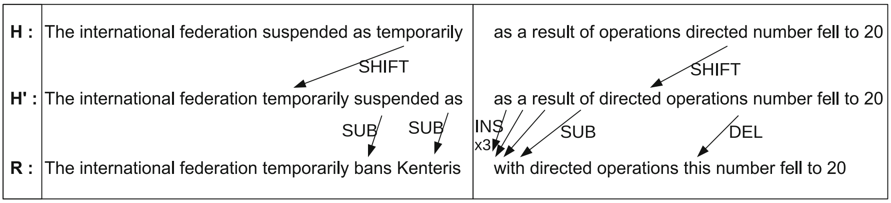
H : The international federation suspended as temporarily
as a result of operations directed number fell to 20
SHIFT
SHIFT
H' : The international federation temporarily suspended as
as a result of directed operations number fell to 20
SUB
SUB
INS
x3
SUB
DEL
R : The international federation temporarily bans Kenteris
with directed operations this number fell to 20

Fig. 9 Examples of TER edits required to turn the translation hypothesis (H) first into a shifted hypothesis (H′) and then into the reference (R) for Arabic→English 1-best ML translation

example, there are two high-posterior 4-grams, 'as a result of' and 'number fell to 20', but only one of them appears in the reference even though the first one would also be correct in a human evaluation.

We note that this analysis is related in spirit to the classification error rate (CER) published earlier (Ueffing and Ney 2007), but at an n-gram level rather than a word-based level. In the CER literature each output word can be assigned a confidence tag, which is then evaluated against the 'correct' tag by means of a Levenshtein distance (Levenshtein 1966) to the reference (using word error rate, WER) or via a bag-of-word approach (using the position-independent error rate, PER). In our case, each hypothesised n-gram can be assigned a confidence tag, which is evaluated against the 'correct' tag obtained by an extended Levenshtein algorithm allowing for shifts (TER).

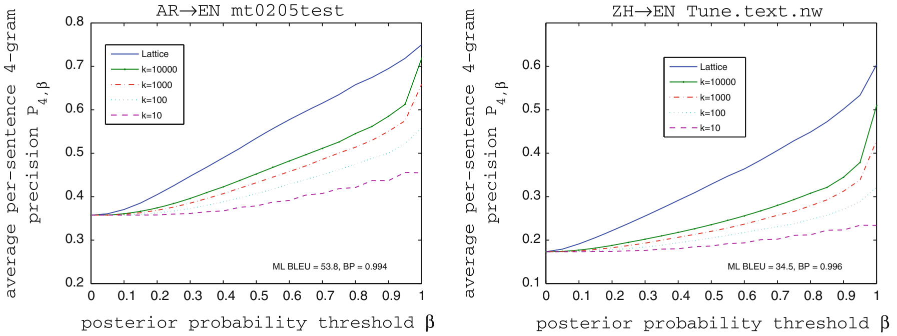

Fig. 10 Average per-sentence 4-gram precisions for Arabic→English (*left*) and Chinese→English (*right*) ML 1-best computed using the full lattice and k-best lists of the specified sizes

7.1.2 Evidence space size and reference precisions

The previous sections report n-gram posteriors of the 1-best translation output, which were computed using a translation lattice that contained potentially millions of alternative translations. One advantage of computing n-gram posteriors from a lattice instead of a k-best list is that a much larger evidence space can be exploited. The larger evidence space of the lattice can improve n-gram posterior probability estimates, leading to improved precision. Arabic→English (left) and Chinese→English (right) 4-gram precisions for a range of posterior probability thresholds are shown in Fig. 10. These posteriors are computed using the full lattice or a k-best list of the specified size. We see that expanding the k-best list size from 1,000 to 10,000 translation hypotheses only slightly improves the precision, whereas much higher precisions are possible if the full evidence space of the lattice is used. Although individual hypotheses beyond the 10,000th in a k-best list might have low posterior probability, their aggregate probability is substantial and useful for accurate estimation of n-gram posterior probability confidence measures. An identical conclusion can be drawn at all n-gram orders, so we do not report these results here.

We can compute the proportion of lattice probability mass missing from a k-best list as the ratio of the sum of posterior probabilities of hypotheses in the k-best list to the sum taken over the full lattice (Blackwood 2010). These statistics can be computed exactly by pushing weights to the final state in the log semiring (Mohri et al. 2008). Figure 11 plots the proportion of lattice probability mass missing from k-best lists of size $k = 1{,}000$ hypotheses (left) and $k = 20{,}000$ hypotheses (right) as a function of the number of lattice n-grams for the Arabic→English mt0205tune set. The $k = 1{,}000$ plot shows there are many sentences for which the top 1,000 hypotheses accounts for only a relatively small proportion of the total lattice probability mass. Comparing $k = 1{,}000$ and $k = 20{,}000$ shows that longer k-best lists do account for a larger proportion of the lattice probability mass. However, there are still a fair number of sentences, particularly the longer sentences, for which $k = 20{,}000$ lists account for less than 50 % of the total lattice probability mass. Table 5 shows the average

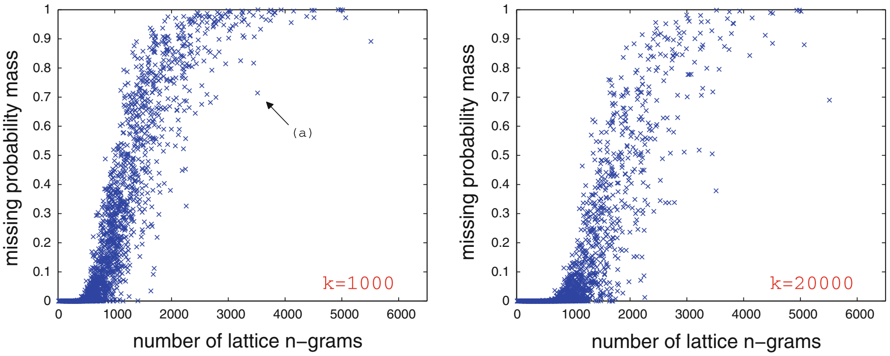

Fig. 11 Proportion of lattice probability mass missing from k-best lists of size $k = 1{,}000$ (*left*) and $k = 20{,}000$ (*right*) versus # of lattice n-grams (Arabic→English mt0205tune). For example, the point labelled (*a*) in the *left-hand plot* corresponds to a lattice containing 3,515 n-grams (for $n = 1 \ldots 4$) with 71.4 % of the lattice probability mass missing from the 1,000-best list

Table 5 Average proportion (%) of missing probability mass by k-best list size for the NIST Arabic→English mt0205tune and mt0205test translation testsets

k	mt0205tune	mt0205test
1,000	24.41	24.91
10,000	13.96	14.27
20,000	11.73	12.00
50,000	9.30	9.52
100,000	7.78	7.98

proportion of missing probability mass in k-best lists of various sizes for mt0205tune and mt0205test.

7.1.3 Single and multiple reference precision experiments

To ensure a good correlation between BLEU and human assessments of MT quality, we should use as many human references as resources allow (Papineni et al. 2002). We compared the effect of using single and multiple references on the precision of high confidence n-grams. Figure 12 shows 4-gram precisions for Arabic→English (left) and Chinese→English (right) computed with respect to each of the individual references, and using the union of the four references. 4-gram precisions with respect to each of the individual references are observed to be very similar. Precision with respect to the union of the references is considerably higher, rising steadily as the posterior probability threshold β is increased. Again, similar trends are observed at all n-gram orders (not reported here). Comparing single versus multiple reference precisions helps us to interpret the WMT and FAUST experiments described in Sects. 7.2 and 7.3, where there are one and two references, respectively. The precision plots of Fig. 12 indicate that single reference confidence measures underestimate the true system performance.

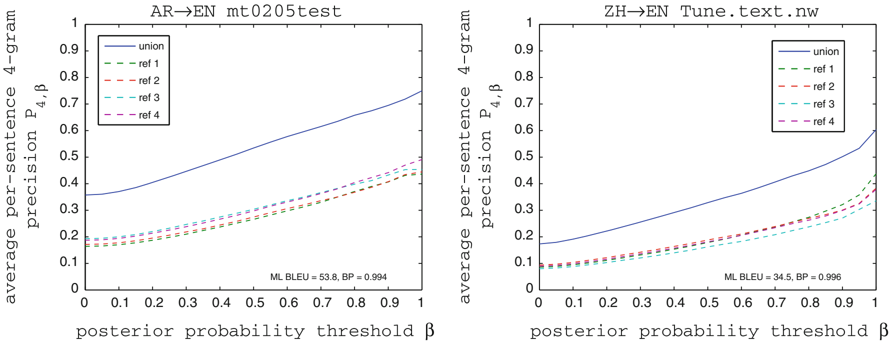

Fig. 12 Average per-sentence 4-gram precisions for Arabic→English (*left*) and Chinese→English (*right*) ML 1-best as a function of β computed with respect to single (*ref 1*, *2*, *3*, *4*) or multiple (*union*) reference translations

the newspaper " constitution " quoted brigadier abdullah krishan **, the chief of police in** karak governorate **(521 km south @-@ west of amman) as saying that the seizure** took place after **police received information that** there were attempts by the group to sell for more than $ 100 thousand dollars , **the police rushed to** the arrest in possession .

Fig. 13 Arabic→English ML 1-best translation hypothesis segmented as a sequence of high and low confidence subsequences. The high confidence subsequences are shown in *bold*

7.1.4 Confidence-based hypothesis segmentation

Posterior probability confidence measures can be used to segment a lattice of translation hypotheses into an alternating sequence of high and low confidence regions (Blackwood 2010; Blackwood et al. 2010b). Figure 13 shows an Arabic→English mt0205tune ML 1-best translation hypothesis that has been segmented according to the posterior probability of its 4-grams. Each 4-gram with posterior probability greater than the confidence threshold $\beta = 0.8$ is considered to be of high-confidence. The high-confidence partial hypotheses in the ML 1-best are therefore formed from consecutive, overlapping high-confidence 4-grams. For this example, the longest high-confidence subsequence is 14 words, consisting of 11 consecutive 4-grams. High-confidence subsequences correspond to partial hypotheses for which there is consensus amongst the translations in the first-pass evidence space. High-confidence subsequences are often of higher quality than low-confidence subsequences. Segmenting translation hypotheses in this way shows how n-gram posterior probability confidence measures can be used to identify low-confidence portions of translation hypotheses that may benefit from re-decoding, post-processing, targeted application of specific models, or user input in an interactive translation setting.

7.2 French→English (WMT 2010)

Lattice MBR decoding of WMT 2010 French→English lattices provides absolute BLEU gains over the ML baseline hypotheses of +0.5, +0.5, and +0.4 for

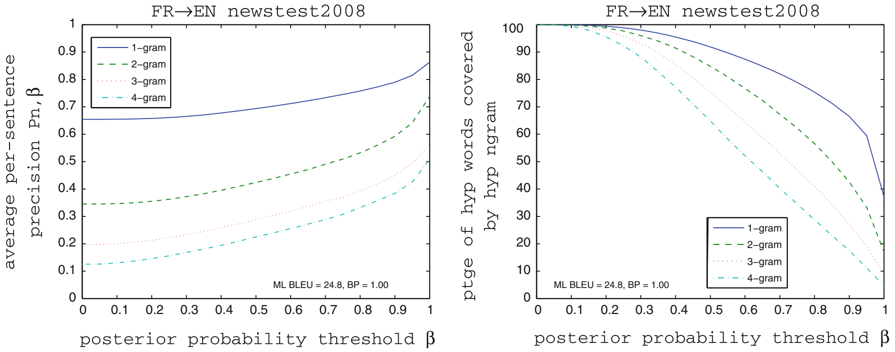

Fig. 14 Average per-sentence n-gram precisions (*left*) and hypothesis coverage (*right*) for WMT 2010 French→English ML translations at a range of posterior probability thresholds $0 \leq \beta \leq 1$

newstest2008, newstest2009, and newstest2010, respectively.[6] This section considers the reliability of n-gram posterior probability confidence measures in French→English translation.

Figure 14 shows reference precisions (left) and hypothesis coverage (right) by order for n-grams in the newstest2008 first-pass ML 1-best translations. Compared to the Arabic→English precisions shown in Fig. 4, these precisions are quite a lot lower. However, the WMT 2010 precisions are computed with respect to a single reference translation. Comparing the single vsersus multiple reference precision plots of Fig. 12 shows that French→English 4-gram precision is quite similar to that of Arabic→English when evaluated using a single reference. These results show that it is possible to identify trusted subsequences of the ML 1-best hypotheses in French→English translation, even if precisions measured against the single reference translation appear quite low.

7.3 FAUST confidence experiments

This section reports lattice MBR decoding performance and precision experiments for the FAUST Spanish→English and English→Spanish translation tasks described in Sect. 5.

7.3.1 Spanish→English

We evaluate lattice MBR decoding performance for translations from the original source sentences 'os', and from cleaned versions of the source data produced by human translators 'cs0' and 'cs1'. Table 6 shows case-insensitive BLEU scores for first-pass HiFST and MBR translations computed with respect to the union of the two available references. We see that when translating the noisy Spanish→English input data, the BLEU score of the ML 1-best is considerably lower than that obtained by translating from the clean data. This suggests that one of the best ways to improve

[6] The baseline ML scores are 24.8, 28.5 and 28.8, respectively. We believe these LMBR gains to be significant given the nature of the task which uses a single English reference.

Table 6 First-pass Spanish→English 1-best translation and lattice MBR BLEU scores for FAUST dev and test sets using noisy or cleaned versions of the source-language input data

	noisy 'os'		clean 'cs0'		clean 'cs1'	
	dev	test	dev	test	dev	test
HiFST	36.3	35.9	47.6	46.9	45.9	45.9
+LMBR	36.2	35.9	48.6	47.9	47.1	46.7

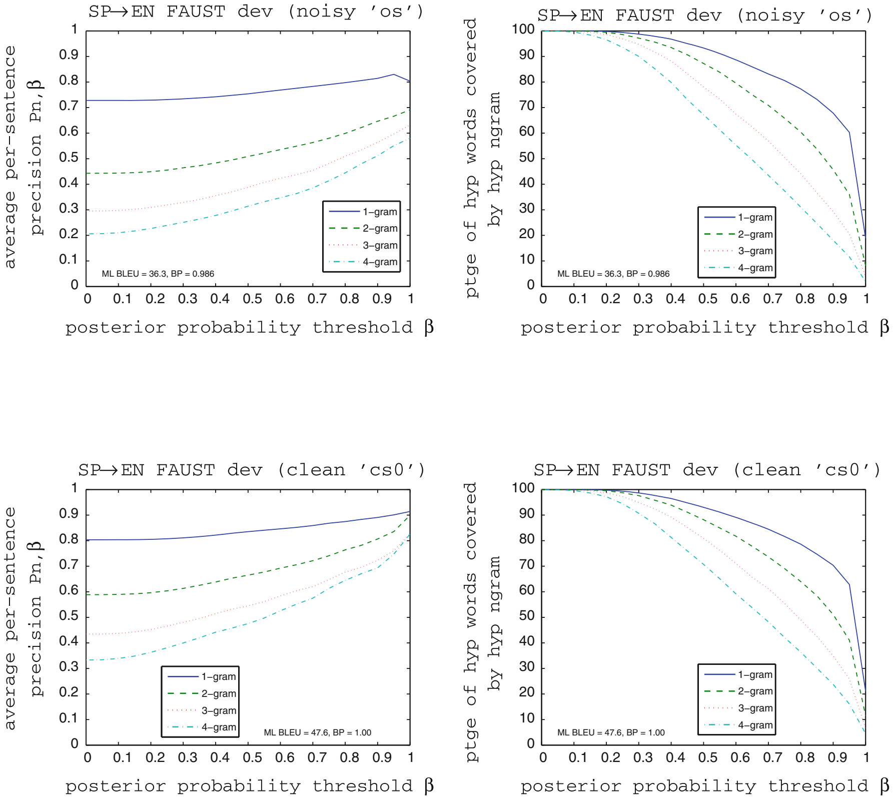

Fig. 15 Average per-sentence *n*-gram precisions (*left*) and hypothesis coverage (*right*) for FAUST Spanish→English ML translations generated from noisy data 'os' (*top*) and cleaned data (*bottom*) at a range of posterior probability thresholds $0 \leq \beta \leq 1$

the quality of noisy translation requests might be to automatically correct possible sources of noise such as spelling mistakes. We note that this is could also be linked to the use of source-language paraphrases, mentioned earlier (Buzek et al. 2010; Resnik et al. 2010). Our results show that lattice MBR over lattices generated from noisy data provides no gain. For the clean 'cs0' and 'cs1' lattices, gains of around +1.0 BLEU are obtained by lattice MBR.

Figure 15 shows per-sentence reference precisions and hypothesis coverage by *n*-gram order for FAUST Spanish→English dev set ML 1-best translations generated

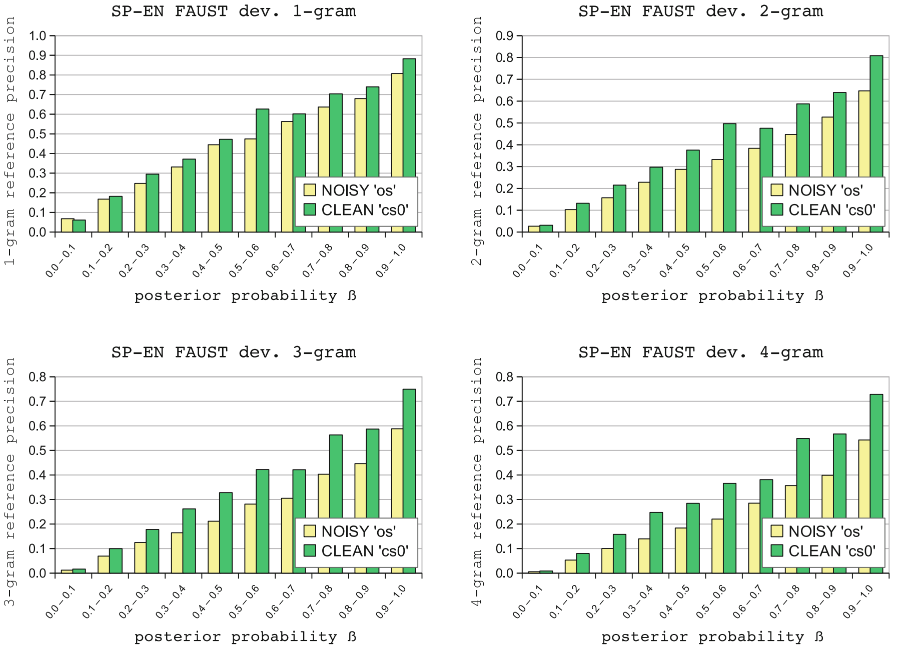

Fig. 16 FAUST Spanish→English dev set reference precisions by n-gram posterior probability computed from translations of the noisy (*top*) and cleaned (*bottom*) source sentences

from the noisy data 'os' and the cleaned version produced by one of the human translators 'cs0'. Comparing precisions and counts shows that at the same confidence threshold $\beta = 0.6$, 55 % of the words output when translating noisy data can be covered by high confidence 4-grams with reference precision 0.35, while these figures rise to 59 % and 0.53 when translating the cleaned source data. These results confirm that cleaning the source input sentences prior to translation leads to greatly improved confidence estimates.

Figure 16 shows how reference precision varies with n-gram posterior probability for n-grams computed from the full lattice when translating the noisy original source 'os' and cleaned source 'cs0' sentences. For both the noisy and cleaned source translations, n-grams with low posterior probability are seen to have very low reference precision. Very low posterior n-grams are seen to indicate low-confidence partial hypotheses. Higher confidence n-grams generally have higher precision. Comparing the precision of high-confidence n-grams u with posterior probability $0.8 \leq p(u|\mathcal{E}) \leq 1.0$ computed from the noisy original source 'os' and cleaned source 'cs0' sentences shows that much higher precisions can be obtained from the cleaned source, particularly as n increases where differences become greater. Removing noise from the input sentences results in higher quality translations (as measured by the BLEU score) and more accurate estimates of n-gram posterior probabilities computed from the lattice.

Table 7 First-pass English→Spanish 1-best translation and lattice MBR BLEU scores for FAUST dev and test sets using noisy or cleaned versions of the source-language input data

	noisy ‘os’		clean ‘cs0’	
	dev	test	dev	test
HiFST	32.0	31.2	35.5	32.8
+LMBR	33.2	33.0	37.3	34.7

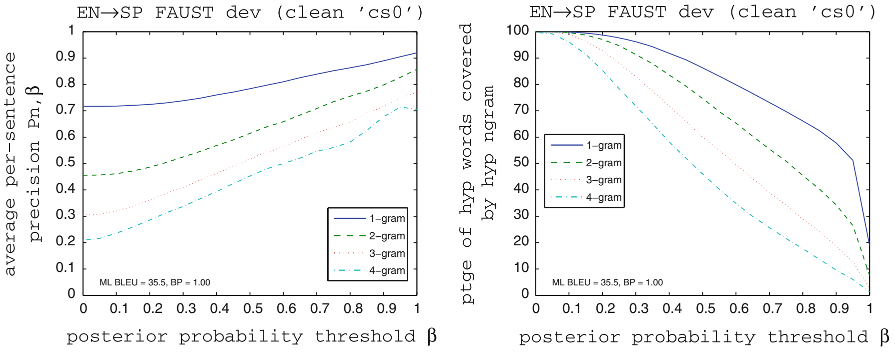

Fig. 17 Average per-sentence *n*-gram precisions (*left*) and hypothesis coverage (*right*) for FAUST English→Spanish ML translations generated from cleaned data ‘cs0’ at a range of posterior probability thresholds $0 \leq \beta \leq 1$

7.3.2 English→Spanish

The results of MBR decoding over English→Spanish lattices generated from the noisy original source ‘os’ and cleaned source data ‘cs0’ are shown in Table 7. For translating from English into Spanish we observe much larger MBR gains than translation in the opposite direction: +1.8 BLEU on the noisy test set, and +1.9 BLEU on the cleaned source test set.

Figure 17 plots reference precisions and hypothesis coverage by *n*-gram order for the dev ML 1-best translations. Both precisions and coverage plots display similar trends to those of Spanish→English translation as the posterior probability threshold β is varied.

7.4 Multi-source translation confidence

Multi-source translation (Och and Ney 2001; Schroeder et al. 2009) is possible whenever the source-language sentence is available in multiple languages. The motivation for multi-source translation is that some of the ambiguity that must be resolved in translating between one pair of languages may not be present in a different pair. In the following experiments, *n*-gram confidence is computed from multiple lattices using the interpolated *n*-gram posterior probabilities of Eqs. (10) and (11) in Sect. 4.4 (Blackwood 2010). Table 8 shows that multi-source MBR combination of WMT 2010

Table 8 Case-insensitive IBM BLEU scores for lattice MBR and two-way multi-source combination of French→English and Spanish→English lattices ($\gamma_{\mathrm{FR}} = 0.55$ and $\gamma_{\mathrm{ES}} = 0.45$)

Configuration	newstest2008	newstest2009	newstest2010
FR→EN			
HiFST+5g	24.8	28.5	28.8
+LMBR	25.3	29.0	29.2
ES→EN			
HiFST+5g	25.2	26.8	30.1
+LMBR	25.4	26.9	30.3
FR→EN + ES→EN			
LMBR	27.2	30.4	32.0

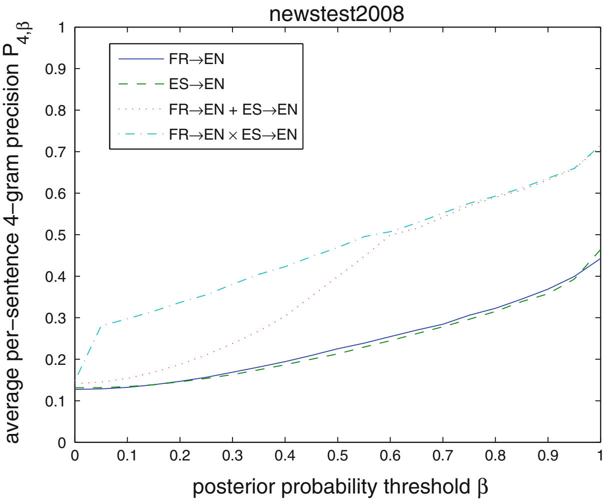

Fig. 18 WMT 2010 FR→EN and ES→EN single-lattice and multi-source 4-gram reference precisions computed using a weighted sum or product of n-gram posterior probabilities

French→English and Spanish→English lattices using a weighted combination of n-gram posteriors leads to very good gains in BLEU score, especially considering that (i) only relatively small gains are observed for single-system lattice MBR, and (ii) these scores are computed with respect to a single reference translation.[7]

We now show that multi-source translation also leads to substantial improvements in the reliability of our n-gram posterior probability confidence measure. Figure 18 shows 4-gram reference precisions for French→English and Spanish→English lattice MBR hypotheses over a range of confidence thresholds $0 \leq \beta \leq 1$. We also provide the multi-source reference precisions for 4-grams computed from the combined evidence space of both lattices using the weighted sum or product. We observe that the reference precisions of n-grams computed from the individual lattices are very similar over the full range of confidence thresholds. Combining the evidence space of

[7] Lattice MBR decoding results reproduced from Pino et al. (2010).

multiple lattices leads to greatly improved 4-gram precisions, with even larger gains than the multi-input hybrid translation confidence experiments reported in de Gispert et al. (2010). For $\beta \geq 0.6$, both combination methods have very similar precisions. For lower confidence thresholds, the weighted product has higher precision. This is because when n-gram confidence is computed as a weighted product of n-gram posterior probabilities, the n-gram must have high posterior probability in each individual lattice in order to have high confidence in the combination.

8 Summary and discussion

We have presented an empirical study of n-gram posterior probability confidence measures for SMT. We first described an efficient and practical algorithm based on a symbol-specific variant of the forward procedure that can be used to compute n-gram posterior probabilities from an MT word lattice. The efficiency of this algorithm is such that it is possible to incorporate n-gram confidence measures computed from large lattices in real-time interactive and computer-aided translation systems.

We used our algorithm to perform a detailed empirical study of n-gram confidence measures for a variety of language pairs and experimental frameworks. We have shown that high posterior probability n-grams are a reliable predictor of whether or not a word sequence is found in the human reference translations, and that such n-grams occur often enough to be useful. This motivates the use of the n-gram posterior probability as a confidence measure. Our confidence experiments have investigated the importance of the evidence space size, the effect of using single or multiple references, compared precisions computed using lattices generated from both noisy and cleaned versions of the source-language input sentences, and shown that multiple-lattice system combination can be used to obtain more reliable estimates of confidence in multi-source translation.

Computing confidence measures at the n-gram level, rather than the word- or sentence-level, allows for many interesting applications. These include improvements to interactive machine translation (Casacuberta et al. 2009) and computer-aided translation (Barrachina et al. 2009), error driven paraphrasing and re-translation (Resnik et al. 2010), and confidence-based lattice segmentation and re-decoding with targeted models in regions of low confidence (Blackwood et al. 2010b).

Acknowledgements This work was supported in part under the GALE program of the Defense Advanced Research Projects Agency, Contract No. HR0011-06-C-0022 and the European Union Seventh Framework Programme (FP7-ICT-2009-4) under Grant Agreement No. 247762.

References

Allauzen C, Riley M, Schalkwyk J, Skut W, Mohri M (2007) OpenFst: a general and efficient weighted finite-state transducer library. In: Proceedings of the ninth international conference on implementation and application of automata (CIAA). Springer lecture notes in computer science, Prague, pp 11–23

Barrachina S, Bender O, Casacuberta F, Civera J, Cubel E, Khadivi S, Lagarda AL, Ney H (2009) Statistical approaches to computer-assisted translation. Comput Linguist 25(1):3–28

Bender O, Matusov E, Hahn S, Hasan S, Khadivi S, Ney H (2007) The RWTH Arabic-to-English spoken language translation system. In: Proceedings of the automatic speech understanding workshop (ASRU), Kyoto, pp 396–401

Blackwood G (2010) Lattice rescoring methods for statistical machine translation. PhD Thesis, University of Cambridge and Clare College, Cambridge

Blackwood G, de Gispert A, Byrne W (2010a) Efficient path counting transducers for minimum Bayes-risk decoding of statistical machine translation lattices. In: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL): short papers, Uppsala, pp 27–32

Blackwood G, de Gispert A, Byrne W (2010b) Fluency constraints for minimum Bayes-risk decoding of statistical machine translation lattices. In: Proceedings of the 23rd international conference on computational linguistics (COLING), Beijing, pp 71–79

Blatz J, Fitzgerald E, Foster G, Gandrabur S, Goutte C, Kulesza A, Sanchis A, Ueffing N (2004) Confidence estimation for machine translation. In: Proceedings of the 20th international conference on computational linguistics (COLING), Geneva, pp 315–321

Brants T, Popat AC, Xu P, Och FJ, Dean J (2007) Large language models in machine translation. In: Proceedings of the joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL), Prague, pp 858–867

Buzek O, Resnik P, Bederson BB (2010) Error driven paraphrase annotation using mechanical turk. In: Proceedings of the NAACL-HLT workshop on creating speech and language data with Amazon's mechanical turk, Los Angeles, pp 217–221

Casacuberta F, Civera J, Cubel E, Lagarda AL, Lapalme G, Macklovitch E, Vidal E (2009) Human interaction for high quality machine translation. Commun ACM 52(10):135–138

Chiang D (2007) Hierarchical phrase-based translation. Comput Linguist 33(2):201–228

Cormen TH, Leiserson CE, Rivest RL, Stein C (2001) Introduction to algorithms, 2nd edn. MIT Press, Cambridge

de Gispert A, Iglesias G, Blackwood G, Banga ER, Byrne W (2010) Hierarchical phrase-based translation with weighted finite-state transducers and shallow-*n* grammars. Computat Linguist 36(3):505–533

DeNero J, Kumar S, Chelba C, Och F (2010) Model combination for machine translation. In: Proceedings of human language technologies: the 11th annual conference of the North American chapter of the Association for Computational Linguistics (HLT-NAACL), Los Angeles, pp 975–983

Deng Y, Byrne W (2008) HMM word and phrase alignment for statistical machine translation. IEEE Trans Audio Speech Lang Process 16(3):494–507

González-Rubio J, Ortiz-Martínez D, Casacuberta F (2010) Balancing user effort and translation error in interactive machine translation via confidence measures. In: Proceedings of the annual meeting of the Association for Computational Linguistics (ACL): short papers, Uppsala, pp 173–177

Graff D, Kong J, Chen K, Maeda K (2007) English gigaword, 3rd edn. Linguistic Data Consortium, Linguistic Data Consortium

Habash N, Rambow O (2005) Arabic tokenization, part-of-speech tagging and morphological disambiguation in one fell swoop. In: Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL), Ann Arbor, pp 573–580

Iglesias G, de Gispert A, Banga ER, Byrne W (2009a) Rule filtering by pattern for efficient hierarchical translation. In: Proceedings of the 12th conference of the European chapter of the Association of Computational Linguistics (EACL), Athens, pp 380–388

Iglesias G, de Gispert AR, Banga E, Byrne W (2009b) Hierarchical phrase-based translation with weighted finite state transducers. In: Proceedings of human language technologies: the 10th annual conference of the North American chapter of the Association for Computational Linguistics (HLT-NAACL), Boulder, pp 433–441

Iglesias G, Allauzen C, Byrne W, de Gispert A, Riley M (2011) Hierarchical phrase-based translation representations. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Edinburgh, pp 1373–1383

Jiang H (2005) Confidence measures for speech recognition: a survey. Speech Commun 45:455–470

Jiang L, Huang X (1998) Vocabulary-independent word confidence measure using subword features. In: Proceedings of the 5th international conference on spoken language processing (ICSLP), vol 7, Sydney, pp 3245–3248

Kneser R, Ney H (1995) Improved backing-off for *m*-gram language modeling. In: Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), vol 1, Detroit, pp 181–184

Kumar S, Byrne W (2003) A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In: Proceedings of human language technologies: the annual conference of the North American chapter of the Association for Computational Linguistics (HLT-NAACL), Edmonton, pp 63–70

Kumar S, Byrne W (2004) Minimum Bayes-risk decoding for statistical machine translation. In: Proceedings of human language technologies: the annual conference of the North American chapter of the Association for Computational Linguistics (HLT-NAACL), Boston, pp 169–176

Levenshtein V (1966) Binary codes capable of correcting deletions, insertions, and reversals. Sov Phys Dokl 10:707–710

Mohri M (1997) Finite-state transducers in language and speech processing. In: Computational linguistics, vol 23. MIT Press, Cambridge, pp 269–311

Mohri M, Pereira F, Riley M (2008) Speech recognition with weighted finite-state transducers. In: Handbook on speech processing and speech communication. Springer, New York

Och FJ (2003) Minimum error rate training in statistical machine translation. In: 41st annual meeting of the Association for Computational Linguistics, proceedings of the conference, Sapporo, pp 160–167

Och FJ, Ney H (2001) Statistical multi-source translation. In: MT summit VIII: machine translation in the information age, proceedings, Santiago de Compostela, pp 253–258

Och FJ, Ney H (2002) Discriminative training and maximum entropy models for statistical machine translation. In: 40th annual meeting of the Association for Computational Linguistics, proceedings of the conference. Philadelphia, pp 295–302

Papineni K, Roukos S, Ward T, Zhu W-J (2002) BLEU: a method for automatic evaluation of machine translation. In: 40th annual meeting of the Association for Computational Linguistics, proceedings of the conference, Philadelphia, pp 311–318

Pino J, Iglesias G, de Gispert A, Blackwood G, Brunning J, Byrne W (2010) The CUED HiFST system for the WMT10 translation shared task. In: Proceedings of the joint fifth workshop on statistical machine translation and MetricsMATR, Uppsala, pp 155–160

Rahim M, Lee C-H, Juang B-H (1997) Discriminative utterance verification for connected digits recognition. IEEE Trans Speech Audio Process 5(3):266–277

Resnik P, Buzek O, Hu C, Kronrod Y, Quinn A, Bederson BB (2010) Improving translation via targeted paraphrasing. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Cambridge, pp 127–137

Rosti A-V, Matsoukas S, Schwartz R (2007) Improved word-level system combination for machine translation. In: Proceedings of the annual meeting of the Association of Computational Linguistics (ACL), Prague, pp 312–319

Schroeder J, Cohn T, Koehn P (2009) Word lattices for multi-source translation. In: Proceedings of the 12th conference of the European chapter of the Association for Computational Linguistics (EACL), Athens, pp 719–727

Sim K-C, Byrne W, Gales M, Sahbi H, Woodland P (2007) Consensus network decoding for statistical machine translation system combination. In: Proceedings of the international conference on acoustics, speech, and signal processing (ICASSP), vol 4, Honolulu, pp 105–108

Snover M, Dorr BJ, Schwartz R, Micciulla L, Makhoul J (2006) A study of translation edit rate with targeted human annotation. In: Proceedings of the 7th biennial conference of the Association for Machine Translation in the Americas (AMTA), Cambridge, pp 223–231

Specia L, Saunders C, Turchi M, Wang Z, Shawe-Taylor J (2009a) Improving the confidence of machine translation quality estimates. In: MT summit XII: proceedings of the twelfth machine translation summit, Ottawa, pp 136–143

Specia L, Turchi M, Cancedda N, Dymetman M, Cristianini N (2009b) Estimating the sentence-level quality of machine translation systems. In: EAMT-2009: proceedings of the 13th annual conference of the European Association for Machine Translation, Barcelona, pp 28–35

Tromble R, Kumar S, Och F, Macherey W (2008) Lattice minimum Bayes-risk decoding for statistical machine translation. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), Honolulu, pp 620–629

Ueffing N, Ney H (2005) Word-level confidence estimation for machine translation using phrase-based translation models. In: Proceedings of the conference on human language technology and empirical methods in natural language processing (HLT-EMNLP), Vancouver, pp 763–770

Ueffing N, Ney H (2007) Word-level confidence estimation for machine translation. Comput Linguists 33(1):9–40

Ueffing N, Och FJ, Ney H (2002) Generation of word graphs in statistical machine translation. In: EMNLP-2002: proceedings of the 2002 conference on empirical methods in natural language processing, Philadelphia, pp 156–163

Wessel F, Schlüter R, Macherey K, Ney H (2001) Confidence measures for large vocabulary continuous speech recognition. IEEE Trans Speech Audio Process 9:288–298