Translating Unknown Words by Analogical Learning

Philippe Langlais and **Alexandre Patry**
Dept. I.R.O.
Université de Montréal
{felipe,patryale}@iro.umontreal.ca

Abstract

Unknown words are a well-known hindrance to natural language applications. In particular, they drastically impact machine translation quality. An easy way out commercial translation systems usually offer their users is the possibility to add unknown words and their translations into a dedicated lexicon. Recently, Stroppa and Yvon (2005) have shown how analogical learning alone deals nicely with morphology in different languages. In this study we show that analogical learning offers as well an elegant and effective solution to the problem of identifying potential translations of unknown words.

1 Introduction

Analogical reasoning has received some attention in cognitive science and artificial intelligence (Gentner et al., 2001). It has been for a long time a faculty assessed in the so-called SAT Reasoning tests used in the application process to colleges and universities in the United States. Turney (2006) has shown that it is possible to compute relational similarities in a corpus in order to solve 56% of typical analogical tests quizzed in SAT exams. The interested reader can find in (Lepage, 2003) a particularly dense treatment of analogy, including a fascinating chapter on the history of the notion of analogy.

The concept of *proportional analogy*, denoted $[A : B = C : D]$, is a relation between four entities which reads: "A is to B as C is to D". Among proportional analogies, we distinguish *formal analogies*, that is, ones that arise at the graphical level, such as [*fournit* : *fleurit* = *fournie* : *fleurie*] in French or [*believer* : *unbelievable* = *doer* : *undoable*] in English. Formal analogies are often good indices for deeper analogies (Stroppa and Yvon, 2005).

Lepage and Denoual (2005) presented the system ALEPH, an intriguing example-based system entirely built on top of an automatic formal analogy solver. This system has achieved state-of-the-art performance on the IWSLT task (Eck and Hori, 2005), despite its striking purity. As a matter of fact, ALEPH requires no distances between examples, nor any threshold.[1] It does not even rely on a tokenization device. One reason for its success probably lies in the specificity of the BTEC corpus: short and simple sentences of a narrow domain. It is doubtful that ALEPH would still behave adequately on broader tasks, such as translating news articles.

Stroppa and Yvon (2005) propose a very helpful algebraic description of a formal analogy and describe the theoretical foundations of *analogical learning* which we will recap shortly. They show both its elegance and efficiency on two morphological analysis tasks for three different languages.

Recently, Moreau et al. (2007) showed that formal analogies of a simple kind (those involving suffixation and/or prefixation) offer an effective way to extend queries for improved information retrieval.

In this study, we show that analogical learning can be used as an effective method for translating unknown words or phrases. We found that our approach has the potential to propose a valid translation for 80% of *ordinary* unknown words, that is, words that are not proper names, compound words, or numerical expressions. Specific solutions have been proposed for those token types (Chen et al., 1998; Al-Onaizan and Knight, 2002; Koehn and Knight, 2003).

The paper is organized as follows. We first recall

[1] Some heuristics are applied for speeding up the system.

Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp. 877–886, Prague, June 2007.

in Section 2 the principle of analogical learning and describe how it can be applied to the task of enriching a bilingual lexicon. In Section 3, we present the corpora we used in our experiments. We evaluate our approach over two translation tasks in Section 4. We discuss related work in Section 5 and give perspectives of our work in Section 6.

2 Analogical Learning

2.1 Principle

Our approach to bilingual lexical enrichment is an instance of analogical learning described in (Stroppa and Yvon, 2005). A learning set $\mathcal{L} = \{L_1, \ldots, L_N\}$ gathers N observations. A set of features computed on an incomplete observation X defines an input space. The inference task consists in predicting the missing features which belong to an output space. We denote $I(X)$ (resp. $O(X)$) the projection of X into the input (resp. output) space. The inference procedure involves three steps:

1. Building $\mathcal{E}_{\mathcal{I}}(X) = \{(A, B, C) \in \mathcal{L}^3 \,|\, [I(A) : I(B) = I(C) : I(X)]\}$, the set of input *stems*[2] of X, that is the set of triplets (A, B, C) which form with X an analogical equation.

2. Building $\mathcal{E}_{\mathcal{O}}(X) = \{Y \,|\, [O(A) : O(B) = O(C) : Y], \forall (A, B, C) \in \mathcal{E}_{\mathcal{I}}(X)\}$ the set of solutions to the analogical equations obtained by projecting the stems of $\mathcal{E}_{\mathcal{I}}(X)$ into the output space.

3. Selecting $O(X)$ among the elements of $\mathcal{E}_{\mathcal{O}}(X)$.

This inference procedure shares similarities with the K-nearest-neighbor (k-NN) approach. In particular, since no model of the training material is being learned, the training corpus needs to be stored in order to be queried. On the contrary to k-NN, however, the search for closest neighbors does not require any distance, but instead relies on relational similarities. This purity has a cost: while in k-NN inference, neighbors can be found in time linear to the training size, in analogical learning, this operation requires a computation time cubic in N, the number of observations. In many applications of interest, including the one we tackle here, this is simply impractical and heuristics must be applied.

The first and second steps of the inference procedure rely on the existence of an analogical solver, which we sketch in the next section. One important thing to note at this stage, is that an analogical equation may have several solutions, some being legitimate word-forms in a given language, others being not. Thus, it is important to select wisely the generated solutions, therefore Step 3. In practice, the inference procedure involves the computation of many analogical equations, and a statistic as simple as the frequency of a solution often suffices to separate good from spurious solutions.

2.2 Analogical Solver

Lepage (1998) proposed an algorithm for computing the solutions of a formal analogical equation [*A* : *B* = *C* : ?]. We implemented a variant of this algorithm which requires to compute two edit-distance tables, one between A and B and one between A and C. Since we are looking for subsequences of B and C not present in A, insertion cost is null. Once this is done, the algorithm synchronizes the alignments defined by the paths of minimum cost in each table. Intuitively, the synchronization of two alignments (one between A and B, and one between A and C) consists in composing in the correct order subsequences of the strings B and C that are not in A. We refer the reader to (Lepage, 1998) for the intricacies of this process which is illustrated in Figure 1 for the analogical equation [*even* : *usual* = *unevenly* : ?]. In this example, there are 681 different paths that align *even* and *usual* (with a cost of 4), and 1 path which aligns *even* with *unevenly* (with a cost of 0). This results in 681 synchronizations which generate 15 different solutions, among which only *unusually* is a legitimate word-form.

In practice, since the number of minimum-cost paths may be exponential in the size of the strings being aligned, we consider the synchronization of a maximum of M best paths in each edit-distance table. The worst-case complexity of our analogical solver is $O([|A| \times (|B| + |C|)] + [M^2 \times (|A| + ins(B, C))])$, where the first term corresponds to the computation of the two edit-distance tables,

[2] In Turney's work (Turney, 2006), a stem designates the first two words of a proportional analogy.

4	4	4	4	4	4	n	4	4	3	3	2	1	0	0	0
3	3	3	3	3	3	e	3	3	3	2	1	0	0	0	0
2	2	2	2	2	2	v	2	2	2	1	0	0	0	0	0
1	1	1	1	1	1	e	1	1	1	0	0	0	0	0	0
0	0	0	0	0	0		0	0	0	0	0	0	0	0	0
l	a	u	s	u	◁		▷	u	n	e	v	e	n	l	y

				e	v	e	n			e	v	e	n		
u	**s**	**u**	**a**				**l**	**u**	**n**	e	v	e	n	**l**	**y**

⇒usua-un-l-ly

e				v		e	n			e	v	e	n		
	u	**s**	**u**		**a**		**l**	**u**	**n**	e	v	e	n	**l**	**y**

⇒un-usu-a-l-ly

Figure 1: The top table reports the edit-distance tables computed between *even* and *usual* (left part), and *even* and *unevenly* (right part). The bottom part of the figure shows 2 of the 681 synchronizations computed while solving the equation [*even* : *usual* = *unevenly* : ?]. The first one corresponds to the path marked in bold italics and leads to a spurious solution; the second leads to a legitimate solution and corresponds to the path shown as squares.

and the second one corresponds to the maximum time needed to synchronize them. $|X|$ denotes the length, counted in characters of the string X, whilst $ins(B, C)$ stands for the number of characters of B and C not belonging to A. Given the typical length of the strings we consider in this study, our solver is quite efficient.[3]

Stroppa and Yvon (2005) described a generalization of this algorithm which can solve a formal analogical equation by composing two finite-state transducers.

2.3 Application to Lexical Enrichment

Analogical inference can be applied to the task of extending an existing bilingual lexicon (or transfer table) with new entries. In this study, we focus on a particular enrichment task: the one of translating valid words or phrases that were not encountered at training time.

A simple example of how our approach translates unknown words is illustrated in Figure 2 for the (unknown) French word *futilité*. In this example, translations is inferred by commuting plural and singular words. The inference process lazily captures the fact that English plural nouns ending in *-ies* usually correspond to singular nouns ending in *-y*.

Step 1	source (French) stems
[*activités* : *activité* = *futilités* : *futilité*]	
[*hostilités* : *hostilité* = *futilités* : *futilité*]	...
Step 2a	projection by lexicon look-up
activités↔*actions*	*hostilité*↔*hostility*
hostilités↔*hostilities*	*activité*↔*action*
futilités↔*trivialities,gimmicks*	...
Step 2b	target (English) resolution
[*actions* : *action* = *gimmicks* : ?]	⇒ gimmick
[*hostilities* : *hostility* = *trivialities* : ?]	⇒ triviality
[*hobbies* : *hobby* = *trivialities* : ?]	⇒ triviality
Step 3	selection of target candidates
$\langle$*triviality*, *2*$\rangle$, $\langle$*gimmick*, *1*$\rangle$, ...	

Figure 2: Illustration of the analogical inference procedure applied to the translation of the unknown French word *futilité*.

Formally, we are given a training corpus $\mathcal{L} = \{\langle S_1, T_1\rangle, \ldots, \langle S_N, T_N\rangle\}$ which consists of a collection of N bilingual lexicon entries $\langle S_i, T_i\rangle$. The input space is in our case the space of possible source words, while the output space is the set of possible target words. We define:

$$\forall X \equiv \langle S, T\rangle,\ I(X) = S \text{ and } O(X) = T$$

Given an unknown source word-form S, Step 1 of the inference process consists in identifying source stems which have S as a solution:[4]

$$\mathcal{E}_{\mathcal{I}}(S) = \{\langle i, j, k\rangle \in [1, N]^3 \mid [S_i : S_j = S_k : S]\}.$$

During Step 2a, each source stem belonging to $\mathcal{E}_{\mathcal{I}}(S)$ is projected form by form into (potentially several) stems in the output space, thanks to an operator $proj$ that will be defined shortly:

$$\mathcal{E}_{\langle i,j,k\rangle}(S) = \{T \mid [U : V = W : T]\} \text{ where } (U, V, W) \in (proj_{\mathcal{L}}(S_i) \times proj_{\mathcal{L}}(S_j) \times proj_{\mathcal{L}}(S_k)).$$

[3] Several thousands of equations solved within one second.

[4] All strings in a stem must be different, otherwise, it can be shown that all source words would be considered.

During Step 2b, each solution to those output stems is collected in $\mathcal{E}_{\mathcal{O}}(S)$ along with its associated frequency:

$$\mathcal{E}_{\mathcal{O}}(S) = \bigcup_{\langle i,j,k\rangle \in \mathcal{E}_{\mathcal{I}}(S)} \mathcal{E}_{\langle i,j,k\rangle}(S).$$

Step 3 selects from $\mathcal{E}_{\mathcal{O}}(S)$ one or several solutions. We use frequency as criteria to sort the generated solutions. The projection mechanism we resort to in this study simply is a lexicon look-up:

$$proj_{\mathcal{L}}(S) = \{T \mid \langle S, T\rangle \in \mathcal{L}\}.$$

There are several situations where this inference procedure will introduce noise. First, both source and target analogical equations can lead to spurious solutions. For instance, [*show* : *showing* = *eating* : ?] will erroneously produce *eatinging*. Second, an error in the original lexicon may introduce as well erroneous target word-forms. For instance, when translating the German word *proklamierung*, by making use of the analogy [*formalisiert* : *formalisierung* = *proklamiert* : *proklamierung*], the English equation [*formalised* : *formalized* = *sets* : ?] will be considered if it happens that *proklamiert*↔*sets* belongs to $\mathcal{L}$; in which case, *zets* will be erroneously produced.

We control noise in several ways. The source word-forms we generate are filtered by imposing that they belong to the input space. We also use a (large) target vocabulary to eliminate spurious target word-forms (see Section 3). More importantly, since we consider many analogical equations when translating a word-form, spurious analogical solutions tend to appear less frequently than ones arising from paradigmatic commutations.

2.4 Practical Considerations

Searching for $\mathcal{E}_{\mathcal{I}}(S)$ is an operation which requires solving a number of (source) analogical equations cubic in the size of the input space. In many settings of interest, including ours, this is simply not practical. We therefore resort to two strategies to reduce computation time. The first one consists in using the analogical equations in a generative mode. Instead of searching through the set of stems $\langle S_i, S_j, S_k\rangle$ that have for solution the unknown source word-form S, we search for all pairs (S_i, S_j) to the solutions of $[S_i : S_j = S :?]$ that are valid word-forms of the input space. Note that this is an exact method which follows from the property (Lepage, 1998):

$$[A : B = C : D] \equiv [B : A = D : C]$$

This leaves us with a quadratic computation time which is still intractable in our case. Therefore, we apply a second strategy which consists in computing the analogical equations $[S_i : S_j = S :?]$ for the only words S_i and S_j close enough to S. More precisely, we enforce that $S_i \in v_\delta(S)$ and that $S_j \in v_\beta(S_i)$ for a neighborhood function $v_\gamma(A)$ of the form:

$$v_\gamma(A) = \{B \mid f(B, A) \leq \gamma\}$$

where f is a distance; we used the edit-distance in this study (Levenshtein, 1966). Note that the second strategy we apply is only a heuristic.

3 Resources

In this work, we are concerned with one concrete problem a machine translation system must face: the one of translating unknown words. We are further focusing on the shared task of the workshop on Statistical Machine Translation, which took place last year (Koehn and Monz, 2006) and consisted in translating Spanish, German, and French texts from and to English. For some reasons, we restricted ourselves to translating only into English. The training material available is coming from the *Europarl* corpus. The test material was divided into two parts.[5] The first one (hereafter called `test-in`) is composed of 2 000 sentences from European parliament debates. The second part (called `test-out`) gathers 1 064 sentences[6] collected from editorials of the Project Syndicate website.[7] The main statistics pertinent to our study are summarized in Table 1.

A rough analysis of the 441 different unknown words encountered in the French test sets reveals that 54 (12%) of them contain at least one digit (years, page numbers, law numbers, etc.), 83 (20%) are proper names, 37 (8%) are compound words, 18 (4%) are foreign words (often Latin or Greek

[5]The participants were not aware of this.

[6]We removed 30 sentences which had encoding problems.

[7]`http://www.project-syndicate.com`

	French		Spanish		German	
`test-`	`in`	`out`	`in`	`out`	`in`	`out`
|unknown|	180	265	233	292	469	599
oov%	0.26	1.22	0.38	1.37	0.84	2.87

Table 1: Number of different (source) test words not seen at training time, and out-of-vocabulary rate expressed as a percentage (oov%).

words), 7 words are acronyms, and 4 are tokenization problems. The 238 other words (54%) are ordinary words.

We considered different lexicons for testing our approach. These lexicons were derived from the training material of the shared task by training with GIZA++ (Och and Ney, 2000) —default settings— two transfer tables (source-to-target and the reverse) that we intersected to remove some noise.

In order to investigate how sensitive our approach is to the amount of training material available, we varied the size of our lexicon $\mathcal{L}_T$ by considering different portions of the training corpus ($T = 5\,000$, 10 000, 100 000, 200 000, and 500 000 pairs of sentences). The lexicon trained on the full training material (688 000 pairs of sentences), called $\mathcal{L}_{ref}$ hereafter, is used for validation purposes. We kept (at most) the 20 best associations of each source word in these lexicons. In practice, because we intersect two models, the average number of translations kept for each source word is lower (see Table 2).

Last, we collected from various target texts (English here) we had at our disposal, a vocabulary set $\mathcal{V}$ gathering 466 439 words, that we used to filter out spurious word-forms generated by our approach.

4 Experiments

4.1 Translating Unknown Words

For the three translation directions (from Spanish, German, and French into English), we applied the analogical reasoning to translate the (non-numerical) source words of the test material, absent from $\mathcal{L}_T$. Examples of translations produced by analogical inference are reported in Figure 3, sorted by decreasing order of times they have been generated.

anti-agricole ◇ *(anti-farm,5) (anti-agricultural,3) (anti-rural,3) (anti-farming,3) (anti-farmer,3)*
fleurie ◇ *(flourishing,5) (flourished,4) (flourish,1)*
futilité ◇ *(trivialities,27) (triviality,14) (futile,9) (meaningless,9) (futility,4) (meaninglessness,4) (superfluous,2) (unwieldy,2) (unnecessary,2) (uselessness,2) (trivially,1)* ***(tie,1)*** *(trivial,1)*
butoir ◇ *(deadline,42) (deadlines,33)* **(blows,1)**
court-circuitant ◇ *(bypassing,13) (bypass,12) (bypassed,5) (bypasses,1)*
xviie ◇ *(xvii,18)* ***(sixteenth,3) (eighteenth,1)***

Figure 3: Candidate translations inferred from $\mathcal{L}_{200\,000}$ and their frequency. The candidates reported are those that have been intersected with $\mathcal{V}$. Translations in bold are clearly erroneous.

4.1.1 Baselines

We devised two baselines against which we compared our approach (hereafter ANALOG). The first one, BASE1, simply proposes as translations the target words in the lexicon $\mathcal{L}_T$ which are the most similar (in the sense of the edit-distance) to the unknown source word. Naturally, this approach is only appropriate for pairs of languages that share many cognates (*i.e.*, *docteur* → *doctor*). The second baseline, BASE2, is more sensible and more closely corresponds to our approach. We first collect a set of source words that are close-enough (according to the edit-distance) to the unknown word. Those source words are then projected into the output space by simple bilingual lexicon look-up. So for instance, the French word *demanda* will be translated into the English word *request* if the French word *demande* is in $\mathcal{L}_T$ and that *request* is one of its sanctioned translations.

Each of these baselines is tested in two variants. The first one ($_{id}$), which allows a direct comparison, proposes as many translations as ANALOG does. The second one ($_{10}$) proposes the first 10 translations of each unknown word.

4.1.2 Automatic Evaluation

Evaluating the quality of translations requires to inspect lists of words each time we want to test a variant of our approach. This cumbersome process not only requires to understand the source language,

$\mathcal{L}_T$	5 000		10 000		50 000		100 000		200 000		500 000	
	p%	r%	p%	r%	p%	r%	p%	r%	p%	r%	p%	r%
	`test-in`											
ANALOG	51.4	30.7	55.3	44.4	58.8	64.3	58.2	65.1	59.4	65.2	30.4	67.6
BASE1_{id}	31.6	30.7	32.3	44.4	24.7	64.3	20.3	65.1	20.9	65.2	8.7	67.6
BASE2_{id}	34.5	30.7	37.1	44.4	39.0	64.3	37.8	65.1	34.4	65.2	56.5	67.6
BASE1_{10}	26.7	100.0	28.3	100.0	23.9	100.0	20.0	100.0	16.6	100.0	11.8	100.0
BASE2_{10}	26.3	100.0	30.8	100.0	29.3	100.0	27.6	100.0	24.9	100.0	55.9	100.0
unk	[3 171 , 9.1]		[2 245 , 7.7]		[754 , 4.0]		[456 , 2.9]		[253 , 2.0]		[34 , 1.2]	
	`test-out`											
ANALOG	52.8	28.9	55.3	42.5	52.9	68.8	54.7	74.6	55.7	81.0	43.3	88.2
BASE1_{id}	28.0	28.9	29.0	42.5	27.3	68.8	23.1	74.6	26.8	81.0	22.7	88.2
BASE2_{id}	32.9	28.9	35.0	42.5	32.5	68.8	35.9	74.6	40.8	81.0	59.1	88.2
BASE1_{10}	24.7	100.0	25.9	100.0	25.1	100.0	20.9	100.0	25.2	100.0	25.0	100.0
BASE2_{10}	21.7	100.0	26.4	100.0	27.2	100.0	29.4	100.0	33.6	100.0	57.9	100.0
unk	[2 270 , 8.2]		[1 701 , 6.9]		[621 , 3.4]		[402 , 2.4]		[226 , 1.8]		[76 , 1.4]	

Table 2: Performance of the different approaches on the French-to-English direction as a function of the number T of pairs of sentences used for training $\mathcal{L}_T$. A pair $[n\ ,\ t]$ in lines labeled by *unk* stands for the number of words to translate, and the average number of their translations in $\mathcal{L}_{ref}$.

but happens to be in practice a delicate task. We therefore decided to resort to an automatic evaluation procedure which relies on $\mathcal{L}_{ref}$, a bilingual lexicon which entries are considered correct.

We translated all the words of $\mathcal{L}_{ref}$ absent from $\mathcal{L}_T$. We evaluated the different approaches by computing *response* and *precision* rates. The response rate is measured as the percentage of words for which we do have at least one translation produced (correct or not). The precision is computed in our case as the percentage of words for which at least one translation is sanctioned by $\mathcal{L}_{ref}$. Note that this way of measuring response and precision is clearly biased toward translation systems that can hypothesize several candidate translations for each word, as statistical systems usually do. The reason of this choice was however guided by a lack of precision of the reference we anticipated, a point we discuss in Section 4.1.3.

The figures for the French-to-English direction are reported in Table 2. We observe that the ratio of unknown words that get a translation by ANALOG is clearly impacted by the size of the lexicon $\mathcal{L}_T$ we use for computing analogies: the larger the better. This was expected since the larger a lexicon is, the higher the number of source analogies that can be made and consequently, the higher the number of analogies that can be projected onto the output space. The precision of ANALOG is rather stable across variants and ranges between 50% to 60%.

The second observation we make is that the baselines perform worse than ANALOG in all but the $\mathcal{L}_{500\,000}$ cases. Since our baselines propose translations to each source word, their response rate is maximum. Their precision, however, is an issue. Expectedly, BASE1 is the worst of the two baselines. If we arbitrarily fix the response rate of BASE2 to the one of ANALOG, the former approach shows a far lower precision (*e.g.*, 34.4 against 59.4 for $\mathcal{L}_{200\,000}$). This not only indicates that analogical learning is handling unknown words better than BASE2, but as well, that a combination of both approaches could potentially yield further improvements.

A last observation concerns the fact that ANALOG performs equally well on the out-domain material. This is very important from a practical point of view and contrasts with some related work we discuss in Section 5.

At first glance, the fact that BASE2 outperforms ANALOG on the larger training size is disappointing. After investigations, we came to the conclusion that this is mainly due to two facts. First, the number of unknown words on which both systems were tested is rather low in this particular case (*e.g.*, 34 for the in-domain corpus). Second, we noticed a deficiency of the reference lexicon $\mathcal{L}_{ref}$ for many of those words. After all, this is not surprising since the words unseen in the 500 000 pairs of training sentences, but encountered in the full training corpus (688 000 pairs) are likely to be observed only a few times, therefore weakening the associations automatically acquired for these entries. We evaluate that a third of the reference translations were wrong in this setting, which clearly raises some doubts on our automatic evaluation procedure in this case.

The performance of ANALOG across the three language pairs are reported in Table 3. We observe a drop of performance of roughly 10% (both in precision and response) for the German-to-English translation direction. This is likely due to the heuristic procedure we apply during the search for stems, which is not especially well suited for handling compound words that are frequent in German.

We observe that for Spanish- and German-to-English translation directions, the precision rate tends to decrease for larger values of T. One explanation for that is that we consider all analogies equally likely in this work, while we clearly noted that some are spurious ones. With larger training material, spurious analogies become more likely.

	French		Spanish		German	
T	p%	r%	p%	r%	p%	r%
5	51.4	30.7	52.8	30.3	49.3	23.1
10	55.3	44.4	52.0	45.2	47.6	33.3
50	58.8	64.3	54.0	66.5	44.6	53.2
100	58.2	65.1	53.9	69.1	45.8	55.6
200	59.4	65.2	46.4	71.8	43.0	59.2

Table 3: Performance across language pairs measured on `test-in`. The number T of pairs of sentences used for training $\mathcal{L}_T$ is reported in thousands.

We measured the impact the translations produced by ANALOG have on a state-of-the-art phrase-based translation engine, which is described in (Patry et al., 2006). For that purpose, we extended a phrase-table with the first translation proposed by ANALOG or BASE2 for each unknown word of the test material. Results in terms of word-error-rate (WER) and BLEU score (Papineni et al., 2002) are reported in Table 4 for those sentences that contain at least one unknown word. Small but consistent improvements are observed for both metrics with ANALOG. This was expected, since the original system simply leaves the unknown words untranslated. What is more surprising is that the BASE2 version slightly underperforms the baseline. The reason is that some unknown words that should appear unmodified in a translation, often get an erroneous translation by BASE2. Forcing BASE2 to propose a translation for the same words for which ANALOG found one, slightly improves the figures (BASE2$_{id}$).

	French		Spanish		German	
	WER	BLEU	WER	BLEU	WER	BLEU
base	61.8	22.74	54.0	27.00	69.9	18.15
+BASE2	61.8	22.72	54.2	26.89	70.3	18.05
+BASE2$_{id}$	61.7	22.81	54.1	27.01	70.1	18.14
+ANALOG	61.6	22.90	53.7	27.27	69.7	18.30
sentences	387		452		814	

Table 4: Translation quality produced by our phrase-based SMT engine (base) with and without the first translation produced by ANALOG, BASE2, or BASE2$_{id}$ for each unknown word.

4.1.3 Manual Evaluation

As we already mentioned, the lexicon used as a reference in our automatic evaluation procedure is not perfect, especially for low frequency words. We further noted that several words receive valid translations that are not sanctioned by $\mathcal{L}_{ref}$. This is for instance the case of the examples in Figure 4, where *circumventing* and *fellow* are arguably legitimate translations of the French words *contournant* and *concitoyen*, respectively. Note that in the second example, the reference translation is in the plural form while the French word is not.

Therefore, we conducted a manual evaluation of the translations produced from $\mathcal{L}_{100\,000}$ by ANALOG and BASE2 on the 127 French words of the corpus `test-in`[8] unknown of $\mathcal{L}_{ref}$. Those are the non-numerical unknown words the participating systems in the shared task had to face in the

[8]We did not notice important differences between `test-in` and `test-out`.

contournant	(*49 candidates*)
ANALOG ◇ (circumventing,55) (undermining,20) (evading,19) (circumvented,17) (overturning,16) (circumvent,15) (circumvention,15) (bypass,13) (evade,13) (skirt,12)	
$\mathcal{L}_{ref}$ ◇ **skirting**, **bypassing**, by-pass, overcoming	

concitoyen	(*24 candidates*)
ANALOG ◇ (citizens,26) (fellow,26) (**fellow-citizens**,26) (people,26) (citizen,23) (fellow-citizen,21) (fellows,5) (peoples,3) (civils,3) (fellowship,2)	
$\mathcal{L}_{ref}$ ◇ **fellow-citizens**	

Figure 4: 10 best ranked candidate translations produced by ANALOG from $\mathcal{L}_{200\,000}$ for two unknown words and their sanctioned translations in $\mathcal{L}_{ref}$. Words in bold are present in both the candidate and the reference lists.

in-domain part of the test material. 75 (60%) of those words received at least one valid translation by ANALOG while only 63 (50%) did by BASE2. Among those words that received (at least) one valid translation, 61 (81%) were ranked first by ANALOG against only 22 (35%) by BASE2. We further observed that among the 52 words that did not receive a valid translation by ANALOG, 38 (73%) did not receive a translation at all. Those untranslated words are mainly proper names (*bush*), foreign words (*munere*), and compound words (*rhénanie-du-nord-westphalie*), for which our approach is not especially well suited.

We conclude from this informal evaluation that 80% of ordinary unknown words received a valid translation in our French-to-English experiment, and that roughly the same percentage had a valid translation proposed in the first place by ANALOG.

4.2 Translating Unknown Phrases

Our approach is not limited to translate solely unknown words, but might serve as well to enrich existing entries in a lexicon. For instance, low-frequency words, often poorly handled by current statistical methods, could receive useful translations. This is illustrated in Figure 5 where we report the best candidates produced by ANALOG for the French word *invitées*, which appears 7 times in the 200 000

invitée	(*61 candidates*)
ANALOG ◇ (invited,135) (requested,92) (called,77) (**urged**,75) (guest,72) (**asked**,47) (request,43) (invites,27) (invite,26) (urge,26)	
$\mathcal{L}_{200\,000}$ ◇ **asked**, generate, **urged**	

Figure 5: 10 best candidates produced by ANALOG for the low-frequency French word *invitées* and its translations in $\mathcal{L}_{200\,000}$.

first pairs of the training corpus. Interestingly, ANALOG produced the candidate *guest* which corresponds to a legitimate meaning of the French word that was absent in the training data.

Because it can treat separators as any other character, ANALOG is not bounded to translate only words. As a proof of concept, we applied analogical reasoning to translate those source sequences of at most 5 words in the test material that contain an unknown word. Since there are many more sequences than there are words, the input space in this experiment is far larger, and we had to resort to a much more aggressive pruning technique to find the stems of the sequences to be translated.

expulsent ◇ (expelling,36) (expel,31) (are expelling,23) (are expel,10)
focaliserai ◇ (focus,10) (focus solely,9) (concentrate all,9) (will focus,9) (will placing,9)
dépasseront ◇ (will exceed,4) (exceed,3) (will be exceed,3) (we go beyond,2) (will be exceeding,2)
non-réussite *de* ◇ (lack of success for,4) (lack of success of,4) (lack of success,4)
que vous ***subissez*** ◇ (you are experiencing,2)

Figure 6: Examples of translations produced by ANALOG where the input (resp. output) space is defined by the set of source (resp. target) word sequences. Words in bold are unknown.

We applied the automatic evaluation procedure described in Section 4.1.2 for the French-to-English translation direction, with a reference lexicon being this time the phrase table acquired on the full training material.[9] The response rate in this experiment is particularly low since only a tenth of the sequences

[9]This model contains 1.5 millions pairs of phrases.

received (at least) a translation by ANALOG. Those are short sequences that contain at most three words, which clearly indicates the limitation of our pruning strategy. Among those sequences that received at least one translation, the precision rate is 55%, which is consistent with the rate we measured while translating words.

Examples of translations are reported in Figure 6. We observe that single words are not contrived anymore to be translated by a single word. This allows to capture *1:n* relations such as *dépasseront*↔*will exceed*, where the future tense of the French word is adequately rendered by the modal *will* in English.

5 Related Work

We are not the first to consider the translation of unknown words or phrases. Several authors have for instance proposed approaches for translating proper names and named entities (Chen et al., 1998; Al-Onaizan and Knight, 2002). Our approach is complementary to those ones.

Recently and more closely related to the approach we described, Callison-Burch et al. (2006) proposed to replace an unknown phrase in a source sentence by a paraphrase. Paraphrases in their work are acquired thanks to a word alignment computed over a large external set of bitexts. One important difference between their work and ours is that our approach does not require additional material.[10] Indeed, they used a rather idealistic set of large, homogeneous bitexts (European parliament debates) to acquire paraphrases from. Therefore we feel our approach is more suited for translating "low density" languages and languages with a rich morphology.

Several authors considered as well the translation of new words by relying on distributional collocational properties computed from a huge non-parallel corpus (Rapp, 1999; Fung and Yee, 1998; Takaaki and Matsuo, 1999; Koehn and Knight, 2002). Even if admittedly non-parallel corpora are easier to acquire than bitexts, this line of work is still heavily dependent on huge external resources.

Most of the analogies made at the word level in our study are capturing morphological information. The use of morphological analysis in (statistical) machine translation has been the focus of several studies, (Nießen, 2002) among the first. Depending on the pairs of languages considered, gains have been reported when the training material is of modest size (Lee, 2004; Popovic and Ney, 2004; Goldwater and McClosky, 2005). Our approach does not require any morphological knowledge of the source, the target, or both languages. Admittedly, several unsupervised morphological induction methodologies have been proposed, *e.g.*, the recent approach in Freitag (2005). In any case, as we have shown, ANALOG is not bounded to treat only words, which we believe to be at our advantage.

[10] We do use a target vocabulary list to filter out spurious analogies, but we believe we could do without. The frequency with which we generate a string could serve to decide upon its legitimacy.

6 Discussion and Future Work

In this paper, we have investigated the appropriateness of analogical learning to handle unknown words in machine translation. On the contrary to several lines of work, our approach does not rely on massive additional resources but capitalizes instead on an information which is inherently pertaining to the language. We measured that roughly 80% of ordinary unknown French words can receive a valid translation into English with our approach.

This work is currently being developed in several directions. First, we are investigating why our approach remains silent for some words or phrases. This will allow us to better characterize the limitations of ANALOG and will hopefully lead us to design a better strategy for identifying the stems of a given word or phrase. Second, we are investigating how a systematic enrichment of a phrase-transfer table will impact a phrase-based statistical machine translation engine. Last, we want to investigate the training of a model that can learn regularities from the analogies we are making. This would relieve us from requiring the training material while translating, and would allow us to compare our approach with other methods proposed for unsupervised morphology acquisition.

Acknowledgement We are grateful to the anonymous reviewers for their useful suggestions and to Pierre Poulin for his fruitful comments. This study has been partially funded by NSERC.

References

Yaser Al-Onaizan and Kevin Knight. 2002. Translating named entities using monolingual and bilingual resources. In *Proc. of the 40th ACL*, pages 400–408, Philadelphia, Pennsylvania, USA.

Chris Callison-Burch, Philipp Koehn, and Miles Osborne. 2006. Improved statistical machine translation using paraphrases. In *Proc. of HLT-NAACL*, pages 17–24, New York City, USA.

Hsin-Hsi Chen, Sheng-Jie Hueng, Yung-Wei Ding, and Shih-Chung Tsai. 1998. Proper name translation in cross-language information retrieval. In *Proc. of the 17th COLING*, pages 232–236, Montreal, Québec, Canada.

Matthias Eck and Chiori Hori. 2005. Overview of the IWSLT 2005 evaluation campaign. In *International Workshop on Spoken Language Translation (IWSLT)*, Pittsburgh, Pennsylvania, USA.

Dayne Freitag. 2005. Morphology induction from term clusters. In *Proc. of the 9th CoNLL*, pages 128–135, Ann Arbor, Michigan, USA.

Pascale Fung and Lo Yuen Yee. 1998. An IR approach for translating new words from nonparallel, comparable texts. In *Proc. of the 36th ACL*, pages 414–420, San Francisco, California, USA.

Dedre Gentner, Keith J. Holyoak, and Boicho N. Konikov. 2001. *The Analogical Mind*. The MIT Press, Cambridge, Massachusetts, USA.

Sharon Goldwater and David McClosky. 2005. Improving statistical MT through morphological analysis. In *Proc. of HLT-EMNLP*, pages 676–683, Vancouver, British Columbia, Canada.

Philipp Koehn and Kevin Knight. 2002. Learning a translation lexicon from monolingual corpora. In *Proc. of the ACL Workshop on Unsupervised Lexical Acquisition*, pages 9–16, Philadelphia, Pennsylvania, USA.

Philipp Koehn and Kevin Knight. 2003. Empirical methods for compound splitting. In *Proc. of the 10th EACL*, pages 187–193, Budapest, Hungary.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between european languages. In *Proc. of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 102–121, New York City, USA.

Young-Suk Lee. 2004. Morphological analysis for statistical machine translation. In *Proc. of HLT-NAACL*, Boston, Massachusetts, USA.

Yves Lepage and Etienne Denoual. 2005. ALEPH: an EBMT system based on the preservation of proportionnal analogies between sentences across languages. In *Proc. of IWSLT*, Pittsburgh, Pennsylvania, USA.

Yves Lepage. 1998. Solving analogies on words: an algorithm. In *Proc. of COLING-ACL*, pages 728–734, Montreal, Québec, Canada.

Yves Lepage. 2003. *De l'analogie rendant compte de la commutation en linguistique*. Ph.D. thesis, Université Joseph Fourier, Grenoble, France.

Vladimir. I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, 6:707–710.

Fabienne Moreau, Vincent Claveau, and Pascale Sébillot. 2007. Automatic morphological query expansion using analogy-based machine learning. In *Proc. of the 29th ECIR*, Roma, Italy.

Sonja Nießen. 2002. *Improving Statistical Machine Translation using Morpho-syntactic Information*. Ph.D. thesis, RWTH, Aachen, Germany.

Franz-Joseph Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proc. of the 38th ACL*, pages 440–447, Hong Kong, China.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proc. of the 40th ACL*, pages 311–318, Philadelphia, Pennsylvania, USA.

Alexandre Patry, Fabrizo Gotti, and Philippe Langlais. 2006. Mood at work: Ramses versus Pharaoh. In *Proc. of the HLT-NAACL Workshop on Statistical Machine Translation*, pages 126–129, New York City, USA.

Maja Popovic and Hermann Ney. 2004. Towards the use of word stems and suffixes for statistical machine translation. In *Proc. of the 4th LREC*, pages 1585–1588, Lisbon, Portugal.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated English and German corpora. In *Proc. of the 37th ACL*, pages 519–526, College Park, Maryland, USA.

Nicolas Stroppa and François Yvon. 2005. An analogical learner for morphological analysis. In *Proc. of the 9th CoNLL*, pages 120–127, Ann Arbor, Michigan, USA.

Tanaka Takaaki and Yoshihiro Matsuo. 1999. Extraction of translation equivalents from non-parallel corpora. In *Proc. of the 8th TMI*, pages 109–119, Chester, England.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(3):379–416, Sept.