ALEPH: an EBMT system based on the preservation of proportional analogies between sentences across languages

Yves Lepage, Etienne Denoual

ATR – Spoken language communication labs
Keihanna gakken tosi
619-0288 Kyoto, Japan
{yves.lepage,etienne.denoual}@atr.jp

Abstract

We designed, implemented and assessed ALEPH, a pure example-based machine translation system. It strictly does not make any use of variables, templates or training, does not have any explicit transfer component, and does not require any preprocessing of the aligned examples. It relies on a specific operation: the resolution of analogical equations, that neutralizes translation divergences in an elegant way. Starting only from theoretical results, a system that is state-of-the-art with the top IWSLT 2004 results could be built in six month time. Evaluated on the *Unrestricted Data* track of IWSLT 2004, our system achieved second place in CE, and third place in JE (with best BLEU for this latter track). For this year's evaluation campaign, the features of the system allowed its immediate application to all possible language pairs in the C-STAR tracks.

1. Introduction

We present a novel example-based machine translation system, ALEPH, which relies on an operation specific to language (proportional analogy). We evaluate this system on the tasks proposed during the previous IWSLT 2004 evaluation campaign in the Japanese-English and Chinese-English *Unrestricted Data* track. This evaluation shows that our system would have positioned itself in the top systems for this track (best BLEU in JE track, second position in CE track).

This study demonstrates that it is possible to implement an EBMT system from scratch in a matter of months (0.5 man/year) and achieve reasonable results. Starting only from the theoretical principles of analogy on strings of characters, building a system that is state-of-the-art with the top IWSLT 2004 results could be achieved in as little as six month time.

In the IWSLT 2005 evaluation campaign, this system competed in all C-STAR tracks. Indeed, an appealing feature of this system is that it requires not training whatsoever: the data are just loaded into memory at startup. As a consequence, it can be applied directly to any language pair for which there is sufficient available data.

Following this introduction, the second section of this paper gives a rationale for choosing one and only one basic operation, proportional analogy, to process any sentence of any language. The third section explains in detail the algorithm used to solve proportional analogies, while the fourth section shows its use as a blackbox function to perform translation. The fifth section recalls the experimental conditions of the *Unrestricted Data* track of IWSLT 2004, gives the configuration details for the ALEPH system and ranks it among the participants of IWSLT 2004. The sixth section gives the results of this year's campaign. The final section discusses the results obtained and future research.

2. Divergences across languages

[1] quotes a study on a sample of $19,000$ sentences between English and Spanish showing that one translation pair in three presents divergences. A typical example is the translation of a Spanish verb into an English preposition.

1: *Atravesó* $_{\mathrm{V}}$		0: *It*
2: *el río* $_{\mathrm{N}}$	$\leftrightarrow$	3: *floated* $_{\mathrm{V}}$
3: *flotando* $_{\mathrm{participle.}}$		1: *across* $_{\mathrm{prep.}}$
		2: *the river* $_{\mathrm{N}}$

Approaches that rely on the word as the unit of processing forget the fact that corresponding pieces of information in different languages are indeed distributed over the entire strings and do not necessarily correspond to complete words. For this reason, the correspondence between words given in the example above is in fact not detailed enough. Actually, the ending *-ó* of the first Spanish word accounts for 3rd person singular past tense. So, not only does *atravesó* correspond to the English preposition *across* in its meaning, but, in addition, it also corresponds to another complete word in English

(the pronoun *it*), plus a portion of yet a third English word (the final ending *-ed* of *floated*).

From the monolingual point of view, trivially, any natural language constitutes a "system" in the Saussurian sense of the term. This **systematicity** appears at best in commutations exhibited by proportional analogies like in the following examples. Obviously, any sentence of any language can be cast into a wide number of such proportional analogies, like the following ones:

They swam in the sea.	:	*They swam across the river.*	::	*It floated in the sea.*	:	*It floated across the river.*
It walks across the street.	:	*It walked across the street.*	::	*It floats across the river.*	:	*It floated across the river.*
He swam.	:	*He floated.*	::	*It swam across the river.*	:	*It floated across the river.*

In [2] we have shown how to automatically build tables (or matrices) to visualize the many proportional analogies that can be found in the same resource around a given sentence: in such tables, each cell contains a sentence, and rectangles of four cells constitute proportional analogies. Such proportional analogies reveal the paradigmatic and syntagmatic variations around a given sentence.

From a bilingual point of view, proportional analogies neutralize translation divergences across languages. They leave the choice for a correct translation to an implicit use of the structure of the target language. The correspondences between the source and the target languages in the proportional analogies are solely and entirely responsible not only for the selection of the correct lemmas, but also for the correct word order. For instance, in the example below:

They swam in the sea.	:	*They swam across the river.*	::	*It floated in the sea.*	:	*It floated across the river.*
↕		↕		↕		↕
Nadaron en el mar.	:	*Atravesaron el rio nadando.*	::	*Flotó en el mar.*	:	*x*

the sole resolution of the analogical equation with **the character as the only unit of processing** is sufficient to produce the exact translation of *It floated across the river*, provided that the three sentence pairs on the left are valid translation pairs. The correct Spanish sentence is therefore: $x =$ *Atravesó el rio flotando*.

This demonstrates that no explicit transfer component is needed in this framework: such proportional analogies, as the two above, do not need to tell which word corresponds to which word, or which syntactic structure corresponds to which syntactic structure. Moreover, there is no requirement at all for a particular word to correspond to any other word.

To summarize, the basic element of the proposed framework is the correspondence between two proportional analogies, the sentences of which are valid translation pairs. It can be visualized by the parallelopiped of Figure 1. We shall now explain in detail the vertical planes (Section 3, Proportional analogies) and the horizontal direction (Section 4, Homomorphism between languages of analogical strings) of such a parallelopiped.

3. Proportional analogies

3.1. Scientific background

Our notion of analogies between sentences, or to be more precise between strings of characters, reaches back as far as Euclid and Aristotle: "*A is to B as C is to D*", postulating identity of types for *A*, *B*, *C*, and *D*. The notion has been put forward in morphology by Apollonius Dyscolus and Varro in the Antiquity. In modern linguistics, Saussure [3, part. III, CHAP. IV] considers *analogical equation*s as a typically synchronic operation by which, given two forms of a given word, and only one form of a second word, the fourth missing form is coined: "*honor* is to *hon*$\bar{o}$*rem* as $\bar{o}$*r*$\bar{a}$*tor* is to $\bar{o}$*r*$\bar{a}$*t*$\bar{o}$*rem*"[1]:

$$\bar{o}r\bar{a}t\bar{o}rem \;\; \bar{o}r\bar{a}tor :: hon\bar{o}rem : x \quad \Rightarrow \quad x = honor$$

That analogy applies also to syntax, which is the foundation of our framework, has been advocated by Hermann Paul [4, p. 110] and Bloomfield [5, p. 275]. More recently, Itkonen and Haukioja [6] showed how to deliver grammatical sentences by application of proportional analogies to structural representations.

3.2. Theoretical aspects

While analogy has been largely mentioned and used in linguistics, algorithmic ways to solve proportional analogies between strings of characters have never been proposed,[2] maybe because the operation seems so misleadingly "intuitive". To our knowledge, we were the first to give an algorithm for the resolution of analogical equations in [8]. It is based on the following formalisation of proportional analogies in terms of edit dis-

[1] Latin: $\bar{o}r\bar{a}tor$ (orator, speaker) and *honor* (honour) nominative singular, $\bar{o}r\bar{a}t\bar{o}rem$ and *hon*$\bar{o}$*rem* accusative singular. In the II century BC, *honor* competed with the etymologically correct *honos*.

[2] Except for Copycat [7, p. 205–265] which adopts an artificial intelligence point of view, of little use for linguistic applications.

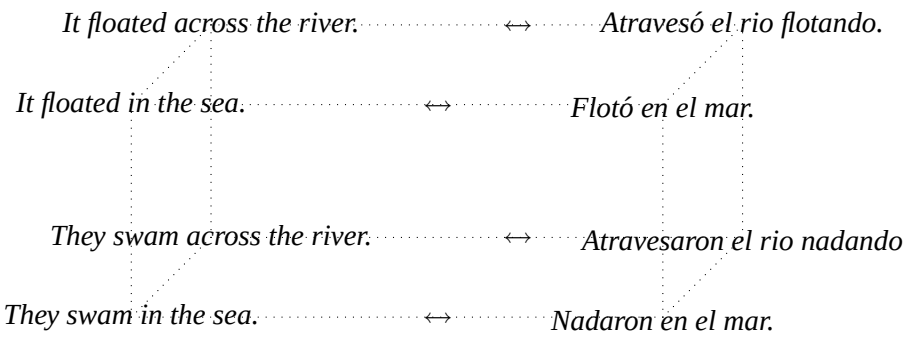

Figure 1: The parallelopiped: four sentences in each language forming a proportional analogy. Four horizontal translation relations exist between the sentences.

tances, or equivalently, in terms of similarity (refer to [9, Chap. 3] for these notions). We denote $\sigma(A, B, \ldots, N)$ as the length of the longest common subsequence in the strings A, B, $\ldots N$, *i.e.*, their similarity. The following formula consistently puts the unknown D on the left of all equal signs, so as to better suit to the resolution of analogical equations.

$A : B :: C : D \quad \Rightarrow$

$$\begin{cases} \sigma(B,D) &= -|A| + |B| + \sigma(A,C) \\ \sigma(C,D) &= -|A| + |C| + \sigma(A,B) \\ \sigma(A,B,C,D) &= -|A| + \sigma(A,B) + \sigma(A,C) \\ |D| &= -|A| + |B| + |C| \end{cases}$$

The step-by-step mechanism we then adopt during resolution is inspired by [6, p. 149], where they take sentence A as the axis against which sentences B and C are compared, and by opposition to which output sentence D is built.

3.3. An example

Rather than explaining once again the algorithm given in [8], we illustrate its application in an actualized way to a particular analogical equation: *like* : *unlike* :: *known* : x. We use words rather than sentences for reasons of space; the same algorithm applies to analogical equations between sentences considered as strings of characters; and it also applies to languages like Japanese, Chinese or Korean where a character is encoded on two bytes instead of just one byte for English.

The similarities between strings A and B, and A and C, are computed in an efficient way using a fast algorithm [10]. Based on a result by [11], only minimal diagonal bands are considered in the matrices. In the following matrices, the algorithm follows the paths noted by values in circles in a way similar to that taken in [12] for the output of an edit distance trace.

e	k	i	l	n	u		k	n	o	w	n
.	.	.	①	⓪	⓪	l	⓪	⓪	.	.	.
.	.	②	1	0	.	i	.	0	⓪	.	.
.	③	2	1	.	.	k	.	.	1	①	.
④	3	2	.	.	.	e	.	.	.	1	①

The succession of moves triggers the copies of characters into the solution D, according to "rules" that tell which character to choose from which string, B or C, according to the moves in both matrices, so that finally, the solution x = *unknown* is output.

dir_{AB}	dir_{AC}	copy onto D	from string
diagonal	diagonal	*n*	C
diagonal	diagonal	*w*	C
diagonal	diagonal	*o*	C
diagonal	diagonal	*n*	C
horizontal	horizontal	*k*	C
horizontal	diagonal	*n*	B
horizontal	diagonal	*u*	B

The example above is simplistic, as it reduces to adding a prefix to *known*. Our algorithm is more powerful as it handles parallel infixing, which is inescapable in the morphology of Semitic languages[3].

aslama : *muslimun* :: *arsala* : $x \quad \Rightarrow \quad x =$ *mursilun* [4]

It is also necessary in our framework because solving analogical equations between sentences involves parallel infixing in the general case. It should be noted that there may be zero, one or several solutions to an analogical equation. Analogical equations are thus a ternary

[3] In particular, if we want to handle the morphology of Arabic in the *Arabic-English C-STAR track* of IWSLT 2005.

[4] Arabic: *arsala* (he sent) and *aslama* (he converted [to Islam]) are verbs 3rd person singular past; *mursilun* (a sender) and *muslimun* (a convert, *i.e.*, a muslim) are agent nouns.

operation, *i.e.*, a mapping $\alpha : \mathcal{L} \times \mathcal{L} \times \mathcal{L} \mapsto \wp(\mathcal{L})$ (with $\mathcal{L}$ the set of strings considered and $\wp(S)$ its power set). The set of the solutions of an analogical equation is:

$$\alpha(A, B, C) = \{D \in \mathcal{L} \mid \ A : B :: C : D \ \}$$

4. Homomorphisms between languages of analogical strings

4.1. Theoretical aspects

Based on proportional analogies, we have shown [13] how to define a family of formal languages, called *languages of analogical strings*. It is important to note that their construction, as is the case with simple contextual[5] grammars [14], does not make any use of nonterminals. Such languages are built by transitive closure starting from a corpus of given sentences (strings of characters) Λ_0. We denote $\alpha(\Lambda, \Lambda, \Lambda)$ as the set of sentences produced by solving all possible analogical equations formed with three sentences in Λ.

$$\alpha(\Lambda, \Lambda, \Lambda) = \{D \mid \exists (A, B, C) \in \Lambda^3, \ A : B :: C : D \ \}$$

Then, the language $\mathcal{L}(\Lambda_0)$ of analogical strings built from a corpus Λ_0 is defined in the following way[6]:

$$\mathcal{L}(\Lambda_0) = \bigcup_{n=0}^{+\infty} \Lambda_n \ \text{ where } \ \Lambda_{n+1} = \alpha(\Lambda_n, \Lambda_n, \Lambda_n)$$

As for the position of such languages in the Chomsky-Schützenberger hierarchy, it is easy to show that the classical regular language $\{a^n | n \geq 1\}$, the context-free language $\{a^n b^n | n \geq 1\}$, and the context-sensitive language $\{a^n b^n c^n | n \geq 1\}$ are all languages of analogical strings. Moreover, we have shown [13] that the famous context-sensitive language $\{a^n b^m c^n d^m \mid m, n \geq 1\}$ used in [15] to refute the context-freeness hypothesis of natural language, is a language of analogical strings. More importantly, every language of analogical strings meets the constant growth property, a property that intervenes partially in the definition of mild context-sensitivity, a notion introduced in [16] to cope with the apparent power of human languages.

The framework for translation by proportional analogies that we propose sees both the source and the target languages as languages of analogical strings that are defined from the set of sentences given in the training corpus. Let us denote $\widehat{A}$ as the (set of) translations of a sentence A. The principle of translation is based on the following intuitive formula that is a transcription of the parallelopiped of Figure 1:

$$A : B :: C : D \ \Leftrightarrow \ \widehat{A} : \widehat{B} :: \widehat{C} : \widehat{D}$$

Using the α operation that structures the source and target languages of analogical strings, an equivalent form of this formula is:

$$\widehat{D} = \widehat{\alpha(A, B, C)} = \alpha(\widehat{A}, \widehat{B}, \widehat{C})$$

This shows that this translation principle "distributes" translation on the arguments of the structuring internal operation α. Thus, it is a homomorphism between two languages of analogical strings that preserves the structuring operation, proportional analogy[7].

For this reason, the translation system described here has been called ALEPH. It is an acronym for Analogy in Languages & Processing by Homomorphism.

4.2. An example

Building on what has been said above, suppose we have a bicorpus at our disposal, *i.e.*, a corpus of aligned sentences in two languages, say, Japanese and English. Suppose that we want to translate the following Japanese input sentence:

濃いコーヒーが飲みたい。[8]

Among all possible pairs of sentences from the bicorpus, we may find the following two Japanese sentences:

紅茶をください。	↔	*May I have some tea, please?*
コーヒーをください。	↔	*May I have a cup of coffee?*

that will allow us to form the following analogical equation:

紅茶をください。 : コーヒーをください。 :: y : 濃いコーヒーが飲みたい。

This equation yields $y =$ 濃い紅茶が飲みたい。[9] If this sentence already belongs to the bicorpus, *i.e.*, if the following translation pair is found in the data:

濃い紅茶が飲みたい。	↔	*I'd like some strong tea, please.*

[5] *Contextual* grammars, not *context-sensitive* grammars!

[6] In fact, $\Lambda_{n+1} \supset \Lambda_n$ because $\quad A : A :: A : x \quad \Rightarrow \quad x = A$.

[7] This is sufficient to solve "difficult" reordering problems. With its translation knowledge reduced to the two translation pairs: $abc \leftrightarrow abc$, $abcabc \leftrightarrow aabbcc$, the system translates members of the regular language $\{ (abc)^n \mid n \in \mathbb{N}^* \}$ into the corresponding members of the context-sensitive language $\{ a^n b^n c^n \mid n \in \mathbb{N}^* \}$, and reciprocally: $(abc)^n \leftrightarrow a^n b^n c^n$, by solving $2 \times (n-2)$ proportional analogies recursively.

[8] Gloss: strong coffee NOMINATIVE-PARTICLE drink-VOLITIVE. Literally: *I want to drink strong coffee.*

[9] Lit.: *I want to drink strong tea.*

then, the following analogical equation is formed with the corresponding English translations:

May I have some tea, please?	:	*May I have a cup of coffee?*	::	*I'd like some strong tea, please.*	: x

By construction, the solution: x = *I'd like a cup of strong coffee.* is a candidate translation of the input sentence: 濃いコーヒーが飲みたい。

5. The core of the ALEPH EBMT system

The following gives the basic outline of our method to perform the translation of an input sentence, using a given bicorpus of aligned sentences:

- Form all analogical equations with the input **sentence** D and with all relevant pairs of **sentences** (A_i, B_i) from the source part of the bicorpus[10];

$$A_i : B_i :: x : D$$

- For those sentences that are solutions of the previous analogical equations and which do not belong to the bicorpus, translate them using the present method recursively. Add them with their newly generated translations to the bicorpus;
- For those sentences $x = C_{i,j}$ that are solutions of the previous analogical equations[11] and which belong to the bicorpus, do the following;
- Form all analogical equations with all possible target language sentences corresponding to the source language sentences[12];

$$\widehat{A_i}^k : \widehat{B_i}^k :: \widehat{C_{i,j}}^k : y$$

- Output the solutions $y = \widehat{D_{i,j}}^k$ of the analogical equations as a translation of D, sorted by frequencies[13].

Although our system has been implemented in the C programming language, the previous algorithm is trivially expressed in Prolog as shown in Figure 2. There are only two predicates: `translation` for translation pairs, and `analogy` to solve analogical equations (with unknown C and $\widehat{D}$ on each respective lines). On the last line, a new translation pair is added to the database of translation facts so that indeed, the system learns as it translates. Also, this program shows that the method is in essence bidirectional.

[10] Relevant pairs of sentences are selected on-the-fly according to a similarity criterion.

[11] One analogical equation may yield several solutions.

[12] Several target sentences may correspond to the same source sentence.

[13] Different analogical equations may yield identical solutions.

% database of translation facts

`translation`(s_1,$\widehat{s_1}$) .
`translation`(s_2,$\widehat{s_2}$) .
⋮
`translation`(s_n,$\widehat{s_n}$) .

% translation routine

`translation`(D,$\widehat{D}$) :–
    `translation`(A,$\widehat{A}$),
    `translation`(B,$\widehat{B}$),
    `analogy`(A,B,C,D),
    `translation`(C,$\widehat{C}$),
    `analogy`($\widehat{A}$,$\widehat{B}$,$\widehat{C}$,$\widehat{D}$),
    `assert`(`translation`(D,$\widehat{D}$)).

Figure 2: A Prolog program for EBMT by proportional analogy. Constants are in lowercase, variables in uppercase.

6. Evaluation and comparison with other systems

We assessed our system on last year's IWSLT task in both Japanese-English and Chinese-English directions in the *Unrestricted Data* track. In this track, no restrictions were imposed on linguistic resources. As for tools, the ALEPH system did not make any use of any NLP tool such as a tagger or the like to preprocess the data. In particular, we chose to place ourselves in the condition of standard natural Japanese and Chinese texts (in which no segmentation appear), so that we had to delete segmentation in the provided test sets! This demonstrates clearly that segmentation is not a necessity to perform a translation task from Japanese or Chinese[14]. As for data, no dictionary was used. The C-STAR corpus of around 160K aligned sentences was used for both language pairs. We refer to this as our "training data", although there is absolutely no training phase within our framework. To avoid the fact that some sentences in the test data may be included in the "training data", we thus assessed our system in two configurations: *standard* and *open*. The difference between the two is that, in the latter, any sentence from the test set was removed from the "training data", if found there.

Some examples of Japanese-English translations are shown in Figure 3. The figures on the left are the frequencies with which each translation candidate has been output[15]. As we assumed that the most frequent candi-

[14] Translation ought to be performed as much as possible on unmodified real texts without preprocessing: we want to evaluate machine translation systems, not preprocessing tools.

[15] Different analogical equations may yield the same solutions (see Section 5).

date should be the most reliable one, the evaluation was performed on the first candidates only.

To summarize the evaluation results obtained with the objective criteria used in this evaluation campaign, the ALEPH system achieves second place in Chinese-English, and third place in Japanese-English. A stand-out point is the achievement in BLEU: a close second for Chinese-English (0.522, first at 0.524), and the best one for Japanese-English (0.634). Unfortunately, we are not in a position to reproduce the subjective evaluation for the translation results output by the ALEPH system.

Table 4: Permitted resources. Copied in part from [17, p. 3]. $\surd$ indicates permitted resources. Our system did not take advantage of any of the permitted resources (this is indicated by $\times$) except for the IWSLT04 corpus and C-STAR aligned sentences.

Resources	Data Track	
	Unrestricted	ALEPH configuration
IWSLT04 corpus	$\surd$	$\surd$
LDC resources	$\surd$	$\times$
tagger	$\surd$	$\times$
chunker	$\surd$	$\times$
parser	$\surd$	$\times$
external bilingual dictionaries	$\surd$	$\times$
other resources	$\surd$	140K additional aligned sentences

7. Evaluation in all IWSLT 2005 C-STAR tracks

The ALEPH system does not require any training phase and is purely bidirectional. These two features made it possible to participate in all C-STAR data tracks of this year's campaign. Only the *open* configuration of the ALEPH system was used this year. Again, as far as data are concerned, we limited ourselves to the use of the core 160,000 C-STAR translation pairs. For the Arabic-English track however, only the 20,000 supplied translation pairs were used, so that, in fact, the ALEPH system will have to be compared with the other systems of the *Supplied* track.

The results are shown in Table 3. Again, for all language pairs, no tool of any sort was used, which means that, prior to translation, no segmentation or tagging whatsoever was performed. Also no dictionary was added to the corpus of example sentences. That is why the results of the ALEPH system should be considered as a sort of baseline for all C-STAR tracks.

8. Discussion and future work

It could have been feared that the algorithmic complexity, which is basically square in the amount of data, would have enormously impaired the method. However, using a simple heuristics to select only relevant pairs entering in analogical equations, allowed us to keep translation times reasonable. Within a time-out of 1 CPU second, the average translation time per sentence was 0.73 second on a 2.8 GHz processor machine with 4 Gb memory.

As we argued here for a pure Example-Based Machine Translation approach, future work should logically focus on achieving a better usage of example data. The present heuristics that is used to select sentences from the corpus in order to form analogical equation is successful only a quarter of the time. Improving the quality of such a heuristics should widen the coverage of the method.

As was mentioned above, in all reported experiments, we did not take advantage of using other resources or NLP tools. The use of dictionaries, paraphrases and the like may improve the performance of the system. We did not take advantage either of an important potential feature of the system: its learning ability. Intermediary translation pairs obtained during the translation of a given test sentence should be useful in the translation of subsequent test sentences.

9. Conclusion

We have shown that the use of a specific operation, namely proportional analogy, may lead to reasonable results in machine translation. Starting only from theoretical results, a system could be built in six month time. Evaluated on the *Unrestricted Data* track of IWSLT 2004, our system achieved second place in CE, and third place in JE (with best BLEU for this latter track). This same system was applied to IWSLT 2005 tasks in all language pairs with similar performance.

The use of an operation that suits by essence the specific nature of linguistic data, *i.e.*, their capacity of commutation on the paradigmatic and syntagmatic axes, allowed us to dispense with any preprocessing of the data whatsoever, an advantage over techniques that require intensive preprocessing. In addition, this operation has the advantage of tackling the issue of divergences between languages in an elegant way: it neutralises them implicitly. As a consequence, the system implemented does not include any explicit transfer component (either lexical or structural).

10. Acknowledgements

This research was supported in part by the National Institute of Information and Commnications Technology.

コーヒーのおかわりをいただけますか。

2318 *I'd like another cup of coffee.*
2296 *May I have another cup of coffee?*
1993 *Another coffee, please.*
1982 *May I trouble you for another cup of coffee?*
1982 *Can I get some more coffee?*
530 *Another cup of coffee, please.*
516 *Another cup of coffee.*
466 *Can I have another cup of coffee?*
337 *May I get some more coffee?*
205 *May I trouble you for another cup of coffee, please?*

小銭をませてください。

924 *Can you include some small change?*
922 *Can you include some small change, please?*
899 *Would you include some small change?*
896 *Include some small change, please.*
895 *I'd like to have smaller bills mixed in.*
895 *Please change this into small money.*
895 *Will you include some small change?*
885 *Could you include some small change, please?*
880 *May I have some small change, too?*

Figure 3: Examples of Japanese-English translations.

11. References

[1] N. HABASH, "Generation-heavy hybrid machine translation," in *Proceedings of the International Natural Language Generation Conference (INLG'02)*, New York, 2002, pp. 185–191.

[2] Y. LEPAGE and G. PERALTA, "Using paradigm tables to generate new utterances similar to those existing in linguistic resources," in *Proceedings of LREC-2004*, vol. 1, Lisbonne, May 2004, pp. 243–246.

[3] F. de SAUSSURE, *Cours de linguistique générale*. Lausanne et Paris: Payot, 1995.

[4] H. PAUL, *Prinzipien der Sprachgeschichte*. Tübingen: Niemayer, 1920.

[5] L. BLOOMFIELD, *Language*. New York: Holt, 1933.

[6] E. ITKONEN and J. HAUKIOJA, *Grammaticalization: Abduction, Analogy, and Rational Explanation*, 1999, pp. 159–175.

[7] D. HOFSTADTER and the Fluid Analogies Research Group, *Fluid Concepts and Creative Analogies*. New York: Basic Books, 1994.

[8] Y. LEPAGE, "Solving analogies on words: an algorithm," in *Proceedings of COLING-ACL'98*, vol. I, Montreal, Aug. 1998, pp. 728–735. [Online]. Available: http://acl.ldc.upenn.edu/P/P98/P98-1120.pdf

[9] G. A. STEPHEN, *String searching algorithms*. Singapore New Jersey London Hong Kong: World scientific, 1994.

[10] L. ALLISON and T. I. DIX, "A bit string longest common subsequence algorithm," *Information Processing Letter*, vol. 23, pp. 305–310, 1986. [Online]. Available: http://www.csse.monash.edu.au/~lloyd/tildeStrings/Align

[11] E. UKKONEN, "Algorithms for approximate string matching," *Information and Control*, vol. 64, pp. 100–118, 1985.

[12] R. A. WAGNER and M. J. FISCHER, "The string-to-string correction problem," *Journal for the Association of Computing Machinery*, vol. 21, no. 1, pp. 168–173, Jan. 1974. [Online]. Available: http://portal.acm.org/citation.cfm?id=321811

[13] Y. LEPAGE, "Analogy and formal languages," in *Proceedings of FG/MOL 2001*, Helsinki, Aug. 2001, pp. 1–12.

[14] L. ILIE, *On Ambiguity in Internal Contextual Languages*. in [**?**], 1998, pp. 29–45.

[15] S. M. SHIEBER, "Evidence against the context-freeness of natural language," *Linguistics and Philosophy*, vol. 8, pp. 333–343, 1985.

[16] A. JOSHI, K. VIJAY-SHANKER, and D. WEIR, *The Convergence of Mildly Context-Sensitive Grammar Formalisms*, 1991, pp. 31–81.

[17] Y. AKIBA, M. FEDERICO, N. KANDO, H. NAKAIWA, M. PAUL, and J. TSUJII, "Overview of the IWSLT04 evaluation campaign," in *Proc. of the International Workshop on Spoken Language Translation*, Kyoto, Japan, 2004, pp. 1–12.

Table 1: Scores for the IWSLT 2004 Chinese-to-English *Unrestricted Data* track: no restriction on linguistic resources. Copied from [17, p. 11].

	mWER	mPER	BLEU	NIST	GTM
[s]System1	**0.379**	**0.319**	**0.524**	**9.56**	**0.748**
[e]ALEPH *standard*	0.434	0.400	0.522	8.42	0.687
[e]ALEPH *open*	0.437	0.404	0.512	8.24	0.682
[s]System2	0.457	0.393	0.440	7.24	0.671
[s]System3	0.525	0.442	0.350	7.36	0.684
[h]System4	0.531	0.427	0.275	7.50	0.666
[s]System5	0.573	0.499	0.243	5.42	0.602
[h]System6	0.578	0.531	0.311	5.92	0.563
[e]System7	0.594	0.487	0.243	6.13	0.611
[r]System8	0.658	0.542	0.162	6.00	0.584
[e]System9	0.846	0.765	0.079	3.64	0.386

Table 2: Scores for the IWSLT 2004 Japanese-to-English *Unrestricted Data* track: no restriction on linguistic resources. Copied from [17, p. 11].

	mWER	mPER	BLEU	NIST	GTM
[h]System1	**0.263**	**0.233**	0.630	10.72	0.796
[s]System2	0.305	0.249	0.619	**11.25**	**0.824**
[e]ALEPH *standard*	0.324	0.300	**0.634**	9.19	0.731
[e]ALEPH *open*	0.437	0.403	0.534	8.97	0.697
[e]System3	0.485	0.420	0.397	7.88	0.672
[r]System4	0.730	0.597	0.132	5.64	0.568

Table 3: Scores for all IWSLT 2005 C-STAR tracks. Unless differently mentioned in *Remarks*, the system (*open* configuration) used the roughly 160,000 translation pairs of the C-STAR corpus in each language pair, and the evaluation was performed with 16 references.

	mWER	mPER	BLEU	NIST	GTM	*Remarks*
Arabic-English	0.527	0.497	0.382	6.22	0.481	20,000 transl.pairs
Korean-English	0.530	0.486	0.412	7.12	0.446	
Chinese-English	0.454	0.418	0.477	7.85	0.553	
Japanese-English	0.361	0.323	0.593	9.82	0.607	
English-Chinese	0.798	0.746	0.098	3.029	0.363	1 reference