Towards automatic acquisition of linguistic features

Yves LEPAGE and **Chooi Ling GOH**[1]
GREYC, University of Caen
F-14032 Caen cedex, France
{yves.lepage,chooiling.goh}@info.unicaen.fr

Abstract

This paper proposes a method to acquire linguistic features from a corpus of short sentences by extracting analogous sentences like *what 's the next station ?* : *where 's the bus station ?* :: *what is the next stop ?* : *where is the bus stop ?* The procedures used to construct clusters of analogous sentences are presented. Experiments performed on roughly 40,000 short sentences from the tourism domain in English and Japanese are reported, and the clusters produced are analyzed and interpreted in terms of linguistic features.

1 Introduction

1.1 Linguistic features as dimensions in a vectorial space

To explain the ultimate goal of the research presented in this paper, let us consider an elementary sentence, like: *Can I have a blanket?* and let us analyze it using standard linguistic terminology. We can say that this sentence is interrogative, that its main verb is *to have*, that the noun *blanket* is singular, etc. Many other linguistic characterizations or features of the sentence or of elements in the sentence can be suggested in this way, and the sum of all these characterizations constitutes an analysis of the sentence.

Any such linguistic characterization, i.e., linguistic feature in the sentence can be seen in opposition to other linguistic features that may be realized to produce a different sentence. For instance, the previous sentence is interrogative by opposition to its affirmative form: *I can have a blanket.* Its main verb could be different, like in: *Can I get a blanket?* The noun *blanket* is singular, in opposition to its plural form: *Can I have blankets?* Etc.

[1]This author is now with ATR-NiCT, Kyoto 619-0288, Japan. New e-mail: chooiling.goh@nict.go.jp.

Thus, the example sentence forms a pair of *analogous sentences* with any sentence that can be produced by changing any of the linguistic features of the sentence. In this way, we have a pair of analogous sentences with the interrogative and affirmative forms: *Can I have a blanket?* : *I can have a blanket.* We also have a pair of analogous sentences when *to have* is exchanged for *to get*: *Can I have a blanket?* : *Can I get a blanket?* And so on.

The final goal of this research is to leverage on large corpora of sentences to automatically perform linguistic analysis, *i.e.*, to characterize any new sentence by its linguistic features. A linguistic feature may be characterized by an example of a pair of sentences, but not any pair of sentences illustrates a linguistic feature. Only if one can find a number of different pairs of analogous sentences can the opposition be thought as reflecting a linguistic feature. For instance, *Can I have a blanket.* : *I can board on the next flight.* does not reflect any linguistic feature, but the following series does.

Can I have a blanket?	:	*I can have a blanket.*
Can I get some small change?	:	*I can get some small change.*
Can I board on the next flight?	:	*I can board on the next flight.*

Such a series of analogous sentences constitutes a dimension in the space of sentences and separates this space into three sub-spaces. The first one contains all sentences similar to the sentences on the left in the series, and the second one contains all sentences similar to the ones on the right. The third sub-space contains all those sentences that are similar to none of the sentences in the series because the opposition expressed by the series is not relevant to them. Figure 1 illustrates this view of a space of sentences in a simple configuration. Three pairs of sentences on each axis

Kristiina Jokinen and Eckhard Bick (Eds.)
NODALIDA 2009 Conference Proceedings, pp. 118–125

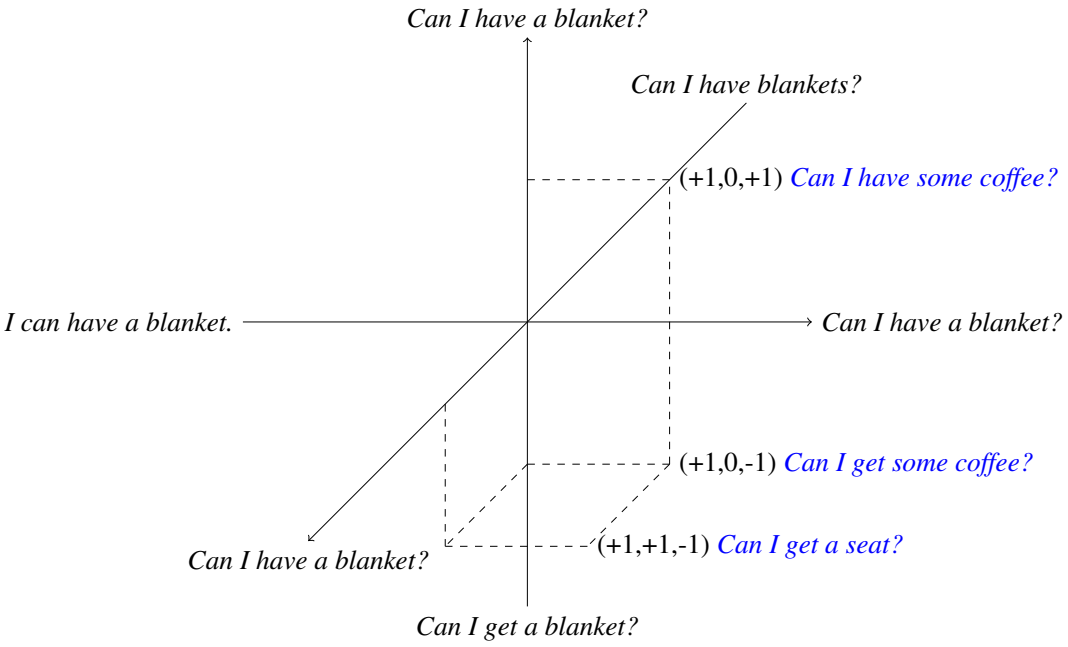

Figure 1: A three-dimensional vectorial space of linguistic features. Each axis stands for the opposition between the two sentences written at both ends.

define three dimensions. Other sentences may be projected in this space according to the possibility for them to enter or not in a series of analogous sentences along any of the dimensions thus taking one of three values: -1 (left), $+1$ (right) or 0 (not relevant) along this dimension.

Such a vectorial space captures those oppositions that are relevant to the sentences of a corpus, thus revealing the linguistic features concealed in that corpus. Such a representation enables the use of any standard vectorial technique for any further desirable computation. The goal of this paper, and the object of the next sections, is not to present such further computations, but to show how it is possible to extract the dimensions defining the space from a corpus of short sentences.

2 Basic Notions

2.1 Analogous Sentences

We follow (Turney, 2006) for the basic notions used in this work:

> Verbal analogies are often written $A : B :: C : D$, meaning *A is to B as C is to D*, for example traffic : street :: water : riverbed.

Following this author, when the relational similarity between two pairs of words is high, we say that the two pairs of words are *analogous*.[2] In this paper, we concentrate on sentences and extend the notion of analogous pairs of words to analogous pairs of sentences. For instance, the two following pairs of sentences are said to be analogous:

Do you have this in darker green? : *Do you have this in dark green?* :: *Smaller, please.* : *Small, please.*

because the relational similarity between the first sentence and the second one is the same as between the third sentence and the fourth one. Logically, following the term *verbal analogies*, we shall call any such two pairs of sentences *sentential analogies*. Here, the relational similarity consists in opposing the positive and comparative forms of two different adjectives: dark : darker :: small : smaller constitute a verbal analogy that sustain the sentential analogy. However, the sole verbal analogy does not imply the sentential analogy because the context in which the words appear constitutes a part of the sentential analogy.

[2] Relational similarity is different from attributional similarity. In this latter case, the correspondence between *attributes* of different words is measured. When this correspondence is high, the two words considered are said to be *synonymous*. In the previous example, water and traffic are not synonymous, clearly showing that relational similarity does not need attributional similarity to exist.

2.2 Series of analogous sentences

When several sentential analogies involve the same pairs of sentences, they form a series of analogous sentences and they can be written on a line like in:

$$A_1 : B_1 :: A_2 : B_2 :: A_3 : B_3 :: \ldots$$

or, in a more convenient way, on a kind of ladder extending over several lines like:

$$\begin{array}{c} A_1 : B_1 \\ A_2 : B_2 \\ A_3 : B_3 \\ \ldots : \ldots \end{array}$$

A requirement would be that, in such a series of analogous sentences, any two pairs of sentences form a sentential analogy. This is the case in the following example where all the three possible sentential analogies hold (see also Table 2):

$$\begin{array}{ll} \textit{Do you have this in darker green?} & : \textit{Do you have this in dark green?} \\ \textit{Smaller, please.} & : \textit{Small, please.} \\ \textit{I'll take the longer one.} & : \textit{I'll take the long one.} \end{array}$$

3 Formalization of Verbal and Sentential analogies

3.1 Previous works on verbal analogies

Measuring the degree of relational similarity between words has received much attention in psychology. Gentner (1983) proposed a model called Structure Mapping Theory (SMT) that has been further elaborated until the present days. Hofstadter and his group have also put forward different proposals, among which the CopyCat model (Hofstadter and the Fluid Analogies Research Group, 1994).

The impact of semantics or pragmatics on verbal analogies may lead to situations where a range of different sources of knowledge may be called upon for the interpretation of specific analogies, leading to quite complex situations like the 'monster analogies' listed by Hoffman (1995). For more standard situations like those found in SAT tests,[3] modern NLP techniques have proved to reach the level of the performance of human beings to identify verbal analogies (Turney and Littman, 2005). Turney (2008) extends and simplifies the previous techniques to propose a uniform approach to synonyms, antonyms, and word associations, through analogies, an approach that could extend to hypernyms/hyponyms, holonyms, *etc.*

Referring to early but fundamental works in linguistics, linguists like de Saussure (1995) or Paul (1920) considered the role of relational similarity, *i.e.*, analogies, in derivational or flexional morphology and even in syntax, from a purely formal point of view. In this way, they justify both the creation of improper, but regular, morphological forms and the production of correct phrasal units.[4] In this trend, we use a definition of analogy between strings of characters that is based on form only, with the risk of capturing meaningless analogies. This formalization is taken from (Lepage, 2004) where the reported measures show that meaningless analogies represent less than 4% of the analogies captured, on the same kind of data that we use in our experiments.

3.2 Measuring relational similarity for sentential analogies

Lepage (2004) measures relational similarity between two pairs of strings (A, B) and (C, D) by verifying the following constraints:

$$\left\{ \begin{array}{rcl} |A|_x - |B|_x & = & |C|_x - |D|_x \\ d(A, B) & = & d(C, D) \end{array} \right.$$

$|A|_x$ is the number of occurrences of character x in string A. d is the canonical edit distance that involves only insertion and deletion with equal weights.[5] As B and C may be exchanged in an analogy, the two constraints above have also to be verified for (A, C) and (B, D). With the previous example, where:

$$\begin{array}{rcl} A & = & \textit{Do you hav}\textbf{e}\textit{ this in dark}\textbf{e}\textit{r gr}\textbf{ee}\textit{n?} \\ B & = & \textit{Do you hav}\textbf{e}\textit{ this in dark gr}\textbf{ee}\textit{n?} \\ C & = & \textit{Small}\textbf{e}\textit{r, pl}\textbf{ea}\textit{s}\textbf{e}\textit{.} \\ D & = & \textit{Small, pl}\textbf{ea}\textit{s}\textbf{e}\textit{.} \end{array}$$

one verifies $d(A, B) = d(C, D) = 2$ and $d(A, C) = d(B, D) = 36$. The relation on the number of occurrences of characters, which is

[3] Scholastic Aptitude Test or Scholastic Assessment Test used in US colleges.

[4] For lack of space, we leave aside the debate about the argument of the poverty of the stimulus (see *The Linguistic Review,* vol. 19, 2003, for arguments and counter-arguments).

[5] This is slightly different from the Levenshtein distance that has substitution as an additional edit operation.

valid for each character, may be illustrated as follows for the character **e**:[6]

$$\begin{array}{ccc} |A|_{\mathbf{e}} - |B|_{\mathbf{e}} & = & |C|_{\mathbf{e}} - |D|_{\mathbf{e}} \\ 4 - 3 & = & 3 - 2 \end{array}$$

The previous characterization of analogies between strings of characters can be expanded in the following way

$$\left\{\begin{array}{rcll} d(A,B) & = & d(C,D) & (i) \\ |A| - |B| & = & |C| - |D| & (ii) \\ |A|_a - |B|_a & = & |C|_a - |D|_a & (iii.a) \\ |A|_b - |B|_b & = & |C|_b - |D|_b & (iii.b) \\ |A|_c - |B|_c & = & |C|_c - |D|_c & (iii.c) \\ |A|_x - |B|_x & = & |C|_x - |D|_x, \forall x & (iv) \end{array}\right.$$

where (ii)–$(iii.c)$ are all logically implied by (iv). $|A|$ denotes the length of A. (ii) expresses the fact that the difference in lengths must be the same for the two pairs of sentences.[7] Conditions $(iii.a)$–$(iii.c)$ are just condition (iv) for three specific characters a, b and c. These three characters are computed over a sample of the sentences of the corpus. They are those characters that exhibit the worst correlations among themselves for all possible values of $|A|_x - |B|_x$. The reason for this is to group pairs of sentences into groups as small as possible.

3.3 Non-transitivity and quality of series of analogous sentences

Notwithstanding, the previous formalization has a deceiving aspect. In this setting, analogy is not a transitive relation, *i.e.*, in the general case, $A : B :: C : D$ and $C : D :: E : F$ do not imply $A : B :: E : F$. An example of such a case is given by the following group of three pairs of sentences:

I prefer the longer one. : *I prefer the long one.*

Do you have this in darker green? : *Do you have this in dark green?*

Smaller, please. : *Small, please.*

where the constraint on distances does not hold between the first and the third pairs of sentences (respective distances 25 and 27).

To compromise with the absence of transitivity when building series of analogous sentences, we shall set a minimal threshold, *i.e.*, the *quality* of a series of pairs of analogous sentences will be defined as the number of actual analogies over the total number of possible analogies. In our experiments, we arbitrarily set this quality level to 90%. We shall refer to series of analogous sentences that exceed this quality level as *analogy clusters*.

4 Automatic Construction of Clusters of Analogous Sentences

4.1 The overall process

In order to automatically build analogy clusters from a corpus of sentences, our method proceeds in several steps:

1. for each sentence of the corpus compute its length and the number of occurrences of the three specific characters. This step is linear in the size of the corpus;

2. for each pair of sentences in the corpus, compute their distance. This step is quadratic in the size of the corpus. Previously sorting the sentences by lengths and imposing $|A| \leq |B|$ reduces the computation by half;

3. for each pair of sentences in a group with the same distance, first compute their difference in lengths and in number of occurrences for the three specific characters and then group pairs of sentences with the same difference in lengths and in number of occurrences of the three specific characters, by applying successive sorts. Distribution sort (or bucket sort) ensures a very fast computation;[8]

4. for each group of pairs of sentences, cluster into analogy clusters by using a greedy method.

4.2 Computing distances between sentences

A very efficient way to compute the distance between two sentences seen as strings of characters is to compute their similarity using the fast bit string algorithm described in (Allison and Dix, 1986) and then derive the value of the canonical

[6] Trivially, $|A|_a - |B|_a = |C|_a - |D|_a \Leftrightarrow |A|_a - |C|_a = |B|_a - |D|_a$.

[7] This property obviously holds because the equality in difference of number of occurrences holds for all the characters in the alphabet.

[8] This is similar in spirit to the technique that consists in building an entire tree-count data-structure as described in (Langlais and Yvon, 2008), but our technique is much more economical as our goal is different and less elaborate.

distance.[9] The above-mentioned algorithm proceeds in two steps, where the first step consists in compiling the first string and the second step computes the similarity. The first step can thus be factored for the computation of the distance between a sentence and all sentences that follow it in increasing lengths, leading to a large speed improvement and to tractable processing time. On a machine with a 2.16 GHz processor, the computation of the distances for 40,000 sentences is achieved in 30 minutes.

4.3 Building analogy clusters

The result of the third step of the process is many groups of pairs of sentences, in which all pairs of sentences share the same distance, the same difference in length and the same difference in number of occurrences for the three specific characters.

Condition (iv) can ultimately be verified between any two pairs of sentences, so as to know whether the analogy holds. For each pair of sentences, the set of other pairs of sentences that form analogies, its analogy set, can be computed and known, so as to know its cardinality.

The clustering process considers the pair of sentences with the largest number of analogies and its analogy set. It successively deletes the pairs of sentences with the least number of analogies from the analogy set until the analogy rate becomes larger than a threshold, 90% in our experiments. The analogy rate is computed as the number of analogies that really exist between all possible pairs of sentences remaining in the analogy set, divided by the square of its cardinality. When the threshold is reached, the cluster is saved and the clustering process proceeds with the next pair of sentences with the largest number of analogies.

5 Experiments

5.1 Corpus used

For experiments, we use an excerpt of the BTEC corpus (Basic Traveling Expressions Corpus). The BTEC corpus is jointly developed by the partners of the C-STAR project.[10] It is a collection of sentences that bilingual travel experts consider useful for people going to or coming from another country. This corpus is widely used in the community of machine translation as it provides translation equivalents in English, Japanese, Chinese, Arabic *etc.*

The excerpt we use is the part that has been released during the international campaign of evaluation of machine translation systems IWSLT 2007 (International Workshop on Spoken Language Translation) (Fordyce, 2007). The following table summarises some statistics about these data.[11]

	English	Japanese
total number of sentences	39,754	36,774
lengths in characters		
shortest sentence:	4	2
longest sentence:	481	234

5.2 Statistics on the clusters produced

The clustering process could build 123,926 English clusters (42,169 for Japanese; in the sequel, the figures in parentheses are for Japanese), of which 118,386 (39,410). contain only two pairs of sentences (called small clusters in Figure 3). The remaining 5,540 (2,759) clusters contain more than 3 pairs of sentences (called large clusters in Figure 3). After distance 40 for the English data and 20 for the Japanese data, large clusters are almost absent. The maximum size of a cluster is 329 (123), obtained with distance 9 (8). Figure 2 plots the sizes of the largest clusters for each distance value.

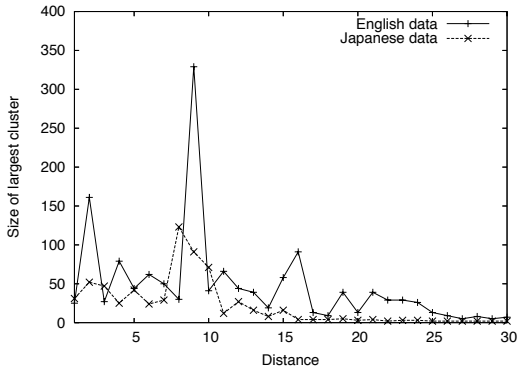

Figure 2: Size of the largest clusters for each distance.

In terms of oppositions, and thus linguistic features, the previous results mean that, for almost

[9] $d(A,B) = |A| + |B| - 2 \times s(A,B)$.

[10] www.c-star.org

[11] As the results presented in the following tables and figures will show, the data at our disposal has been preprocessed to separate punctuations from the preceding words (e.g. *what's* becomes *what 's*) and all words have been lowercased. In reality, this is not necessary for the present experiment, as the method processes the sentences in characters and not in words.

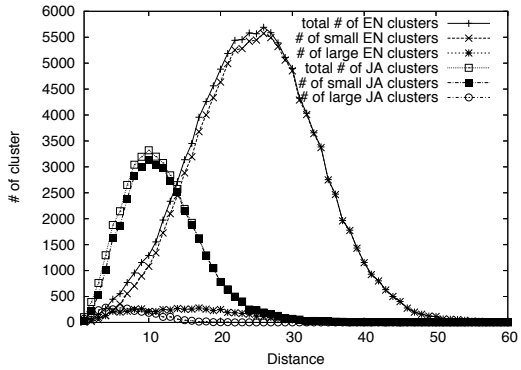

Figure 3: Number of clusters built for each distance.

40,000 English sentences, three times more oppositions could be found that are present in at least two pairs of sentences. However, only 5,500 oppositions are present in more than two pairs of sentences. This leads to a vectorial space of around 5,500 useful dimensions for this corpus.

6 Analysis of the Clusters Produced

In this section, we report on the English data only. Similar trends and explanations can be formulated for the Japanese data.

The largest cluster in our experiment contains 329 pairs of sentences. The interpretation of each cluster has to be made by looking at the opposition between the sentences on the left and the sentences on the right. In this cluster, the pairs of sentences are opposed by the deletion of the ending phrase *, please .* In terms of linguistic feature, one can say that the opposition lies between a neutral and a more polite form of expression. The size of the cluster reflects the optional character of this ending phrase, as one could expect in a corpus that heavily contains expressions of requests.

The next largest cluster contains 161 pairs of sentences. It shows the colloquial use of the contracted form *'s* in place of *is*. One can thus speak about a language level linguistic feature (colloquial vs formal). Again, this is natural in a corpus that necessarily contains traits of oral language.

The third largest cluster contains 91 pairs of sentences. It illustrates the possibility of anteposing *please* at the beginning of a sentence as in: *help me , please .* : *please help me .*

Table 1 shows an example of a cluster where the sentences on the left have the same meaning as the sentences on the right, *i.e.*, they are paraphrases. The linguistic interpretation of this cluster is that the undefinite article *a* can be dropped in certain contexts, especially when expressing a request (sentences ending with: *, please .*)

Table 2 shows another example of a cluster containing sentences with very similar meaning that show that the phrase *where is the* can be substituted for *is there a*.

Other clusters exhibit similar phenonena. Affirmative sentences introduced by *i 'd like to*, are opposed with interrogative sentences introduced by *can i* ended with a circumstancial *here ?*. This may be seen as a structural transformation for near paraphrasing.

Tables 4 and 5 are clusters in which places (*subway station* and *youth hostel*) or predicates (*keep this baggage* and *draw me a map*) are exchanged in similar situational or illocutionary contexts. Such examples, where left and right sentences are not paraphrases, very frequent with smaller clusters, contradicts the impression of paraphrases that one could get by looking too fastly at larger clusters only (see also the remark at the end of Subsection 2.1 and the footnote there). These kinds of clusters do not reflect an opposition in linguistic features but rather show instantiations of semantic features that would be noted like LOC or PRED.

Other clusters make clear some orthographical variations, like the optional use of an hyphen in compound words *check-out*, *take-out etc.* or English vs American writing (*colour* vs *color*), thus reflecting a dialect feature.

Many pairs of sentences appearing in smaller clusters of higher distances appear also in larger clusters with a lower distance. For example, the two pairs of sentences below form one of the small clusters (containing only one sentential analogy).

can i borrow an iron ? : *can i have a blanket ?*
may i borrow an iron ? : *may i have a blanket ?*

But they also appear in a different configuration in a cluster that contains 79 pairs of sentences.

can i borrow an iron ? : *may i borrow an iron ?*
can i have a blanket ? : *may i have a blanket ?*
⋮ : ⋮

The first cluster with only one sentential analogy shows the commutation of the phrase *an iron* with the phrase *a blanket* in a limited context, whereas the second cluster shows the commutation of the two modal verbs *can* and *may*.

# of sent. nlgs	Pairs of sentences		
12	i think there 's a mistake in the bill .	:	i think there 's mistake in the bill .
12	a collect call to japan , please .	:	collect call to japan , please .
12	i 'd like a room with a shower .	:	i 'd like a room with shower .
12	i 'll have a whiskey , please .	:	i 'll have whiskey , please .
11	i 'd like a room with a bath .	:	i 'd like a room with bath .
11	is this a train for chicago ?	:	is this train for chicago ?
13	a one -way ticket , please .	:	one -way ticket , please .
13	a table for two , please .	:	table for two , please .
13	is it a direct flight ?	:	is it direct flight ?
13	i 've got a backache .	:	i 've got backache .
11	porter , please .	:	a porter , please .
11	receipt , please .	:	a receipt , please .
11	i 'm a diabetic .	:	i 'm diabetic .

Table 1: A cluster that illustrates the possible deletion of the undefinite article *a* in some context. One can form only 159 analogies among the 13×13 possibilities. The analogy rate of the cluster is thus: $155/(13 \times 13) = 91.72\%$.

# of sent. nlgs	Pairs of sentences		
11	where is the main area for restaurants ?	:	is there a main area for restaurants ?
11	where is the department store ?	:	is there a department store ?
11	where is the duty -free shop ?	:	is there a duty -free shop ?
11	where is the changing room ?	:	is there a changing room ?
11	where is the sleeping car ?	:	is there a sleeping car ?
11	where is the barber shop ?	:	is there a barber shop ?
11	where is the dining car ?	:	is there a dining car ?
11	where is the restaurant ?	:	is there a restaurant ?
11	where is the gift shop ?	:	is there a gift shop ?
11	where is the telephone ?	:	is there a telephone ?
11	where is the pharmacy ?	:	is there a pharmacy ?

Table 2: A cluster that illustrates a substitution pattern of *where is the* with *is there a*. Its analogy rate is 100%.

# of sent. nlgs	Pairs of sentences		
10	i 'd like to cash this traveler 's check .	:	can i cash this traveler 's check here ?
10	i 'd like to make a hotel reservation .	:	can i make a hotel reservation here ?
10	i 'd like to make a reservation .	:	can i make a reservation here ?
10	i 'd like to check my baggage .	:	can i check my baggage here ?
10	i 'd like to leave my baggage .	:	can i leave my baggage here ?
10	i 'd like to leave my luggage .	:	can i leave my luggage here ?
10	i 'd like to reserve a room .	:	can i reserve a room here ?
10	i 'd like to have dinner .	:	can i have dinner here ?
10	i 'd like to check in .	:	can i check in here ?
10	i 'd like to swim .	:	can i swim here ?

Table 3: A cluster that illustrates the structural transformation of *i 'd like to* ... into *can i ... here ?*

# of sent. nlgs	Pairs of sentences		
4	is there a subway station around here ?	:	is there a youth hostel around here ?
4	how can i get to the subway station ?	:	how can i get to the youth hostel ?
4	is there a subway station near here ?	:	is there a youth hostel near here ?
4	is there a subway station nearby ?	:	is there a youth hostel nearby ?

Table 4: A cluster that examplifies the exchange of place names: *subway station* vs *youth hostel*.

# of sent. nlgs	Pairs of sentences		
4	could you keep this baggage ?	:	could you draw me a map ?
4	keep this baggage , please .	:	draw me a map , please .
4	will you keep this baggage ?	:	will you draw me a map ?
4	please keep this baggage .	:	please draw me a map .

Table 5: A cluster that examplifies the exchange of predicates: *keep this baggage* vs *draw me a map*.

In terms of vectorial space, this confirms the fact that the same sentence may be characterized along several dimensions.

7 Conclusion

We have presented a method that clusters analogous sentences from a corpus of short sentences and helps highlight the linguistic features concealed in a corpus. Such clusters of analogous sentences allow us to build a vectorial space associated with the sentences of a corpus. In an experiment on a corpus of 40,000 English sentences in the tourism domain, we could automatically collect more than 5,000 significant dimensions that represent linguistic oppositions or features. The ones observed on our data extend over a range of linguistic phenomena:

- orthographical variations;
- fronting of interjections;
- exchange of place names, document names, item names etc.;
- normal vs comparative forms of adjectives;
- structural transformations like interrogative vs affirmative;
- exchange of predicates in the same grammatical subject and object context;
- questions in different levels of politeness;
- etc.

References

Lloyd Allison and Trevor I. Dix. 1986. A bit string longest common subsequence algorithm. *Information Processing Letter*, 23:305–310.

Ferdinand de Saussure. 1995. *Cours de linguistique générale*. Payot, Lausanne et Paris, (1st ed. 1916).

Cameron Shaw Fordyce. 2007. Overview of the iwslt 2007 evaluation campaign. In *Proceedings of IWSLT 2007 (International Workshop on Spoken Machine Translation)*, pages 1–12, Trento.

Dedre Gentner. 1983. Structure mapping: A theoretical model for analogy. *Cognitive Science*, 7(2):155–170.

Robert R. Hoffman. 1995. Monster analogies. *AI Magazine*, 11:11–35.

Douglas Hofstadter and the Fluid Analogies Research Group. 1994. *Fluid Concepts and Creative Analogies*. Basic Books, New York.

Philippe Langlais and François Yvon. 2008. Scaling up analogical learning. In *Proceedings of the22nd International Conference on Computational Linguistics (COLING 2008)*, pages 51–54, Manchester, August.

Yves Lepage. 2004. Analogy and formal languages. *Electronic Notes in Theoretical Computer Science*, 53:180–191.

Hermann Paul. 1920. *Prinzipien der Sprachgeschichte*. Niemayer, Tübingen, (1st ed. 1880).

Peter D. Turney and M.L. Littman. 2005. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1–3):251–278.

Peter D. Turney. 2006. Similarity of semantic relations. *Computational Linguistics*, 32(2):379–416.

Peter Turney. 2008. A uniform approach to analogies, synonyms, antonyms, and associations. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 905–912, Manchester, UK, August. Coling 2008 Organizing Committee.

ISSN 1736-6305 Vol. 4
http://hdl.handle.net/10062/9206