Clustering Sentences — Making Sense of Synonymous Sentences

Jussi Karlgren, Björn Gambäck
and Christer Samuelsson
Stockholm

Abstract

The paper describes an experiment on a set of translated sentences obtained from a large group of informants. We discuss the question of transfer equivalence, noting that several target-language translations of a given source-language sentence will be more or less equivalent. Different equivalence classes should form clusters in the set of translated sentences. The main topic of the paper is to examine how these clusters can be found: we consider — and discard as inappropriate — several different methods of examining the sentence set, including traditional syntactic analysis, finding the most likely translation with statistical methods, and simple string distance measures.

1 Introduction

The idea that there is a one-to-one correspondence between sentences in one language and sentences in another is obviously ridiculous to anyone who has tried to translate between any pair of languages. When translating, the aim is not to find **the** correct translation but **a** correct one. For almost any sentence in a source language several sentences in the target language will do: there will not be one good sentence but a set of them, more or less synonymous or **homeosemous** (H. Karlgren, 1974). What a translator (or an information retrieval intermediary) tries to do is to produce a transfer equivalence, i.e., a sentence or a sequence of sentences with a similar or identical pragmatic effect.

This is a decision problem when translating, and an evaluation problem when done. As will be shown below, even for trivially simple source language sentences and utterances there will be a large number of corresponding target language sentences. It would be useful to find a simple method of ranking sentences in such a set to use when evaluating the translation produced by a machine translation (MT) system.

Historically there has been little emphasis on evaluation in the machine translation community, and although that is now starting to change, the methods proposed are often quite *ad hoc*. The strategy chosen for a particular evaluation of course depends on the reasons for the evaluation; or more specifically on who the evaluator is. Developers of MT-systems, end-users and prospective buyers will by necessity evaluate systems in

Proceedings of NODALIDA 1993, pages 143-154

different ways. Following for example Way (1991) MT evaluation strategies are divided into three broad classes:

Typological Evaluation is a developer-oriented strategy aiming at specifying which particular linguistic constructions the system handles satisfactorily and which it does not.

Declarative Evaluation is the strategy commonly used when assessing human translators work; scoring the output with respect to various quality dimensions (such as accuracy, intelligibility and style).

Operational Evaluation is the way end-users and MT-system buyers normally evaluate the systems: measuring how cost- and time-effective a particular system is when used in a specific translation environment.

The principal tool for typological evaluation is a **test suite**, a set of sentences which individually represent specified constructions and hence constitute performance probes. Most work on MT-system evaluation has been concerned with how such a test-suite should be composed, e.g. (King & Falkedal, 1990, and Gambäck *et al*, 1991a, 1991b); however, the methods outlined in this paper follow the declarative evaluation track. Previous methods along this path have normally been "hand-crafted", or based on existing (labour-intensive) methods for the evaluation of human translators' work (Balkan, 1991). Both Thompson (1991) and Su *et al* (1992) have independently worked on automating the process. They present methods for evaluating translation quality based on statistical measurements of a candidate translation against a standard set using simple string-matching algorithms, i.e., ideas quite akin to the ones below.

The rest of the paper is outlined as follows: in the section following we describe an experiment with obtaining a set of translated sentences from a large group of informants. In section 3 we discuss what conclusions can be drawn from the experiment, the key questions being what the structure of the sentence set is and if the set contains clusters. The main topic of the section is how clusters can be found: we consider several different methods of examining the sentence set, including traditional syntactic analysis, finding the most likely translation with statistical methods, and simple string distance measures. Section 4, finally, sums up the previous discussion and points to other possible research directions.

2 Empirical Evidence

In order to find out the extent of divergence of translations, the sentence space, we distributed twelve randomly chosen sentences from a corpus of 4021 spoken English sentences to 1100 Swedish computer scientists. We received 73 answers. The translations were inspected by a professional Swedish translator, and all but a few were considered quite acceptable in a situation corresponding to the one in which they were given. The sentences distributed are shown in table 1 below. They were all in the air traffic information domain, or ATIS, the corpus used by the US government to evaluate the performance of different spoken language understanding systems (Boisen & Bates, 1992).

TABLE 1: Sentences distributed

1	Atlanta to Oakland Thursday.
2	Give me flights from Denver to Baltimore.
3	Which companies fly between Boston and Oakland.
4	Show me all flights from Pittsburgh to Dallas.
5	Show me the names of airlines in Atlanta.
6	What's the cheapest flight from Atlanta to Baltimore.
7	I want to fly from Baltimore to Dallas round trip.
8	Show all flights and fares from Denver to San Francisco.
9	List round trip flights between Boston and Oakland using T W A.
10	What are the flights from Dallas to Boston for the next day.
11	And the ground what is the ground transportation available in the city of Philadelphia.
12	I need a flight leaving Pittsburgh next Monday arriving in Fort Worth before ten a m.

Even the simplest sentence in the test set proved surprisingly divergent: number 1 was translated to twelve different Swedish sentences. For number 12, and the longest sentence in the test set, we received 68 different translations, all of them judged as "good" by the professional translator. Table 2 sums up how the sentences as a whole were translated.

TABLE 2: Summary of responses

Sentence	translations	good	different	most common
1	73	72	12	27
2	74	72	61	4
3	68	66	19	39
4	69	67	36	7
5	73	68	43	9
6	70	68	37	7
7	72	65	27	25
8	70	65	50	10
9	71	71	12	27
10	70	70	62	3
11	68	55	66	2
12	68	68	68	1

A natural choice for a goal translation is to pick the most common translation. For the first sentence in the test set this would give us an appropriate result, the most common translation occurring 27 times; however, for more elaborate sentences this cannot always be done, as shown by sentence number 12. To pick the most typical one, we need to rank the translations. Tables 3 and 4 in the appendix show such frequency ranking for sentences 1 and 5, respectively.

3 What does the study mean?

So, for seventy informants, we received up to seventy non-pathological translations of non-pathological sentences. The question is what the structure of the sentence set is. Are all the sentences synonymous, or does the divergence reflect polysemy on the sentence level? If the sentence set is synonymous, are the sentences just variations over a homogenous space, or are the discernible strategies on some level that can be identified? In effect, what we are asking is if the sentence set contains clusters, or are equidistant, as in figure 1.

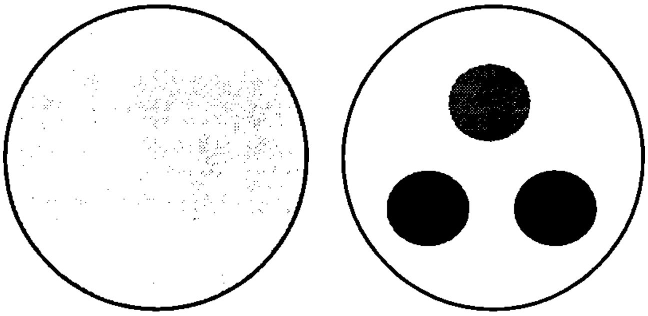


FIGURE 1: Two sentence sets, with equidistant sentences resp. clusters

We will in the following sections consider several different methods of examining the sentence set to find clusters or strategies. First we examine finding the most likely translation with statistical methods, then simple string distance measures, before moving on to traditional syntactic analysis. In passing, we first note that a methodological question that needs to be addressed in a study of this type is whether there is a correct answer to be found as regards the structure of sentence sets. One way of doing this is to ask test subjects to group sentences manually. We have not done this in this small study, but trusted our own judgment as to the likeness between sentences.

3.1 The most likely translation

One obvious way of picking the most typical candidate translation is to choose the most likely one. This is done by comparing the probabilities of the candidate strings. In order to do this, we need a probabilistic language model, i.e., a method of assigning a probability to each string. A simple, but very successful, probabilistic langugage model is the bigram model. In the general case, the probability of a word string $w_1,\ldots,w_n$ is calculated recursively:

$$p(w_1,\ldots,w_n) = p(w_n \mid w_1,\ldots,w_{n-1}) \ldots p(w_1,\ldots,w_{n-1}) =$$

$= \prod_{k=2 \to n} p(w_k \mid w_1,\ldots,w_{k-1})$

The bigram model approximates the factors $p(w_k \mid w_1,\ldots,w_{k-1})$ with the factors $p(w_k \mid w_{k-1})$ — only the word immediately preceeding the current word is taken into account, while the rest of the preceeding string is discarded. Thus, to calculate the string probability all that is used is the probability of each word given any predecessor (bigrams are treated in more detail by e.g. Jelinek, 1990) This gives us the bigram approximation of the string probability of the word string $w_1,\ldots,w_n$:

$$p(w_1,\ldots,w_n) \approx \prod_{k=2 \to n} p(w_k \mid w_{k-1})$$

The probabilites $p(w_k \mid w_{k-1})$ are calculated from the relative frequencies of word pairs in the set of candidate translations corresponding to a sentence in the source language:

$$p(w_k \mid w_{k-1}) \approx f(w_{k-1},w_k) / f(w_{k-1})$$

A different set of probabilities is derived for each source sentence using only the various candidate translations. After all, we are trying to find the most likely translation of this particular sentence. Instead of comparing the probabilities directly, we compare their logarithms, the logarithm function being monotonously increasing. Multiplying probabilities amounts to the same thing as adding their logarithms. Thus

$$\ln\{p(w_1,\ldots,w_n)\} \approx \Sigma_{k=2 \to n} \ln\{p(w_k \mid w_{k-1})\}$$

In order not to penalize longer word strings, the sum is normalized by the string length, giving us the following norm $|w_1,\ldots,w_n|$ of the string $w_1,\ldots,w_n$.

$$|w_1,\ldots,w_n| = - 1 / n \cdot \Sigma_{k=2 \to n} \ln\{p(w_k \mid w_{k-1})\}$$

The minus sign is included to make the norm positive and give more likely sentences smaller norms. This means that $\exp\{- |w_1,\ldots,w_n|\}$ is the geometric mean of the probability of each word w_k in its context, or in other words, its likelihood of occurrence. The probability of a word string $w_1 \ldots w_n$:

$$p(w_1 \ldots w_n) = p(w_n \mid w_1 \ldots w_{n-1}) \cdot p(w_1 \ldots w_{n-1}) =$$
$$= \prod_{k=2 \to n} p(w_k \mid w_1 \ldots w_{k-1}) \approx \prod_{k=2 \to n} p(w_k \mid w_{k-1})$$

Noting that both w_1 and w_n are sentence delimiters (eos), the probability of the sentence *"Atlanta to Oakland Thursday"* is

$$p(\text{eos,Atlanta,to,Oakland,Thursday,eos}) \approx$$
$$\approx p(\text{Atlanta} \mid \text{eos}) \cdot p(\text{to} \mid \text{Atlanta}) \cdot p(\text{Oakland} \mid \text{to}) \cdot$$
$$\cdot\, p(\text{Thursday} \mid \text{Oakland}) \cdot p(\text{eos} \mid \text{Thursday})$$

For the simpler sentence, the bigram statistics produce a similar ranking as do the simple counts of occurrence — not very surprising. Table 3 of the appendix show the bigram rankings for source sentence 1 together with the likelihood (frequency) of the translated target sentences. Table 5 shows that the bigram rankings manage to separate the different translations of sentence 12, a sentence for which pure frequency measures gave no information at all.

3.2 String Distance Methods

Simple string distance measures are designed to match strings of characters rather than strings of words; however, they can be modified to fit these measures as well. Wagner & Fischer (1974) and Lowrance & Wagner (1975) define string distance measures based on primitive string correction operations: replace, delete, insert, and swap. If there is a sequence of edit operations to construct A from B, and N_R, N_D, N_I and N_S are the number of replacements, deletions, insertions and swaps needed in this sequence to convert A to B, and W_R, W_D, W_I, and W_S are costs associated with the operations respectively, the cost of constructing B from A will be the minimum of the following function:

$$D(A,B) = N_R \cdot W_R + N_D \cdot W_D + N_I \cdot W_I + N_S \cdot W_S$$

The distance from string A to string B is defined as the cost of the least cost edit sequence. The measurements were applied to the words as they appeared in the text giving edit distances both character by character and word by word.

After computing the distances between sentences, we need to examine which one of the strings is the most typical. There are standard methods for this type of analysis: we use agglomerative hierarchical clustering, i.e., we assume the sentences all are in separate clusters and repeatedly join the closest pair of clusters until we only have one cluster left. We calculated distance between clusters using two strategies: **complete linkage** and **single linkage** as illustrated in figures 2 and 3.

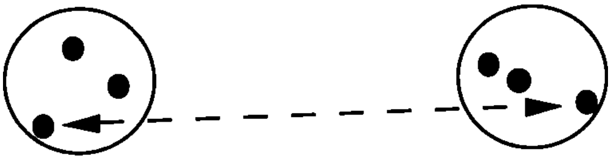


FIGURE 2: Distance between clusters using complete linkage measures

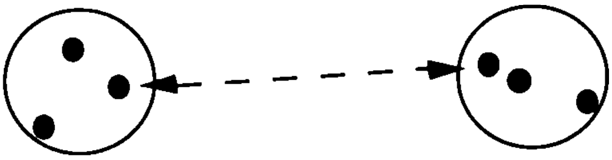


FIGURE 3: Distance between clusters using single linkage measures

In the first case the closest pair of clusters is defined as those where the distance between **furthest** neighbours is minimized, and in the second case as clusters where the distance between **closest** neighbours is minimized. We found that complete linkage gave us a faster clustering, using less steps, and that single linkage yielded a larger number of derivational steps. For most of the experiments, a large number of steps provided more information, so we used single linkage as the preferred strategy.

The results are displayed in the dendrograms in figure 4 in the appendix below (with translations numbered as in tables 3 and 4). In the single linkage based dendrogram for sentence 5 (at the right of figure 4), the two closely related sentences

"visa alla bolag representerade i atlanta"

"vilka flygbolag finns representerade i atlanta"

(translations 21 and 23) are shown to be in different clusters, which naturally is not the desired result.

3.3 Traditional Syntactic Analysis

Consider the following sentence and its translations:

Show me the names of airlines in atlanta.

Vilka flygbolag finns i Atlanta
Vilka flygbolag flyger på Atlanta
Vilka flygbolag trafikerar Atlanta

The three translations correspond to two different syntactic types, and two different propositional contents, whatever way their meaning is analyzed. However, the division by syntactic criteria is different from the division by semantic criteria. Syntax is not the right analysis level to examine complete sentences, since it is concerned with intra-clausal relations, which tend to lose their relevance when larger discourse segments are examined (J. Karlgren, 1993). The aim is to find a level of description with an adequate granularity.

4 Simple Methods: How and Why They Fail

Both statistical and word identity metrics only utilize local information on relatively scarce data. While these types of method are simple to implement, they give relatively little of use for the level of processing we are interested in. Syntactic analysis does not help immediately, as shown by the examples in section 3.3 above. One way to alleviate the arbitrariness of the analysis would be to enlarge the classes of objects studied, by both lexically based methods that equate classes of words — synonym classes, or near synonym classes.

Another way to condense the data better would be to use "demi-structural methods", which add some structure to the text by constructing surface constituents of a relevant level, like complete NP:s and PP:s to perform the analysis. With the advent of reliable surface syntax analysis components (as the ones of, e.g., Voutilainen & Tapanainen, 1993), this could be done with relative little trouble. The idea of leaving certain troublesome grammatical properties to the top level, to be handled by rules of a different type rather than resolving all on the bottom seems to be fruitful.

References

Balkan, Lorna. 1991. *Quality Criteria for MT.* Technical Report. University of Essex, Colchester, England.

Boisen, Sean and Madeleine Bates. 1992. *A Practical Methodology for the Evaluation of Spoken Language Systems.* pp. 162–169, *Proceedings of the 3RD Conference on Applied Natural Language Processing.* Trento, Italy.

Gambäck, Björn, Hiyan Alshawi, Manny Rayner, and David Carter. 1991a. *Measuring Compositionality of Transfer . 1st Meeting of the International Working Group on Evaluation of Machine Translation Systems.* Les Rasses, Switzerland.

Gambäck, Björn, Hiyan Alshawi, Manny Rayner, and David Carter. 1991b. *Measuring Compositionality in Transfer-Based Machine Translation Systems . The ACL Workshop for Evaluation of Natural Language Processing Systems.* University of California, Berkeley, California.

Jelinek, Fred. 1990. *Self-Organizing Language Models for Speech Recognition.* pp. 450–506, READINGS IN SPEECH RECOGNITION. Morgan Kaufmann, San Mateo, California.

Myers, Eugene W. 1986. *An O(ND) Difference Algorithm and its Variations.* pp. 251–266, ALGORITHMICA, 1. Springer-Verlag, New York, New York.

Karlgren, Hans (ed.). 1974. *Homeosemi.* KVAL PUBLIKATION 1974:1. Skriptor, Stockholm, Sweden.

Karlgren, Jussi. 1993. *Syntax in Information Retrieval.* In *Proceedings from the 1st NorFa Doctoral Symposium on Computational Linguistics.* Department of General and Applied Linguistics, Copenhagen University, Copenhagen, Denmark.

King, Maggie and Kirsten Falkedal. 1990. *Using Test Suites in Evaluation of Machine Translation Systems.* pp. 211–216, In *Proceedings of the 13th International Conference on Computational Linguistics,* Volume 2. Helsinki, Finland.

Lowrance, Roy and Robert A. Wagner. 1975. *An Extension of the String-to-String Correction Problem*. pp. 177–183, JOURNAL OF THE ACM, Volume 22, Number 2.

Su, Keh-Yih, Ming-Wen Wu, and Jing-Shin Chang. 1992. *A New Quantitative Quality Measure for Machine Translation Systems*. pp. 433–439, In *Proceedings of the 14th International Conference on Computational Linguistics*. Nantes, France.

Thompson, Henry S. 1991. *Automatic Evaluation of Translation Quality Using Standard Sets. 1st Meeting of the International Working Group on Evaluation of Machine Translation Systems*. Les Rasses, Switzerland.

Voutilainen, Atro and Pasi Tapanainen. 1993. *Ambiguity Resolution in a Reductionistic Parser*. In *The 6th Conference of the European Chapter of the Association for Computational Linguistics*, Utrecht, Holland.

Wagner, Robert A. and M. J. Fischer. 1974. *The String-to-String Correction Problem*. pp. 168–173, JOURNAL OF THE ACM, Volume 21, Number 1.

Way, Andrew. 1991. *Developer Oriented Evaluation of MT Systems*. TECHNICAL REPORT, University of Essex, Colchester, England.

Appendix

TABLE 3: Translation frequencies and bigram probabilities for 1: *"Atlanta to Oakland Thursday"*.

0	27	0,86	atlanta till oakland på torsdag
1	18	0,79	från atlanta till oakland på torsdag
2	12	0,66	atlanta till oakland torsdag
3	1	0,62	från atlanta till oakland torsdag
4	4	0,53	atlanta oakland på torsdag
5	1	0,45	från atlanta till oakland på torsdagen
6	2	0,45	från atlanta till oakland torsdagar
7	1	0,40	atlanta till oakland
8	2	0,34	atlanta oakland torsdag
9	1	0,25	atlanta oakland kommande torsdag
10	1	0,19	på torsdag från atlanta till oakland
11	1	0,10	torsdag atlanta till oakland

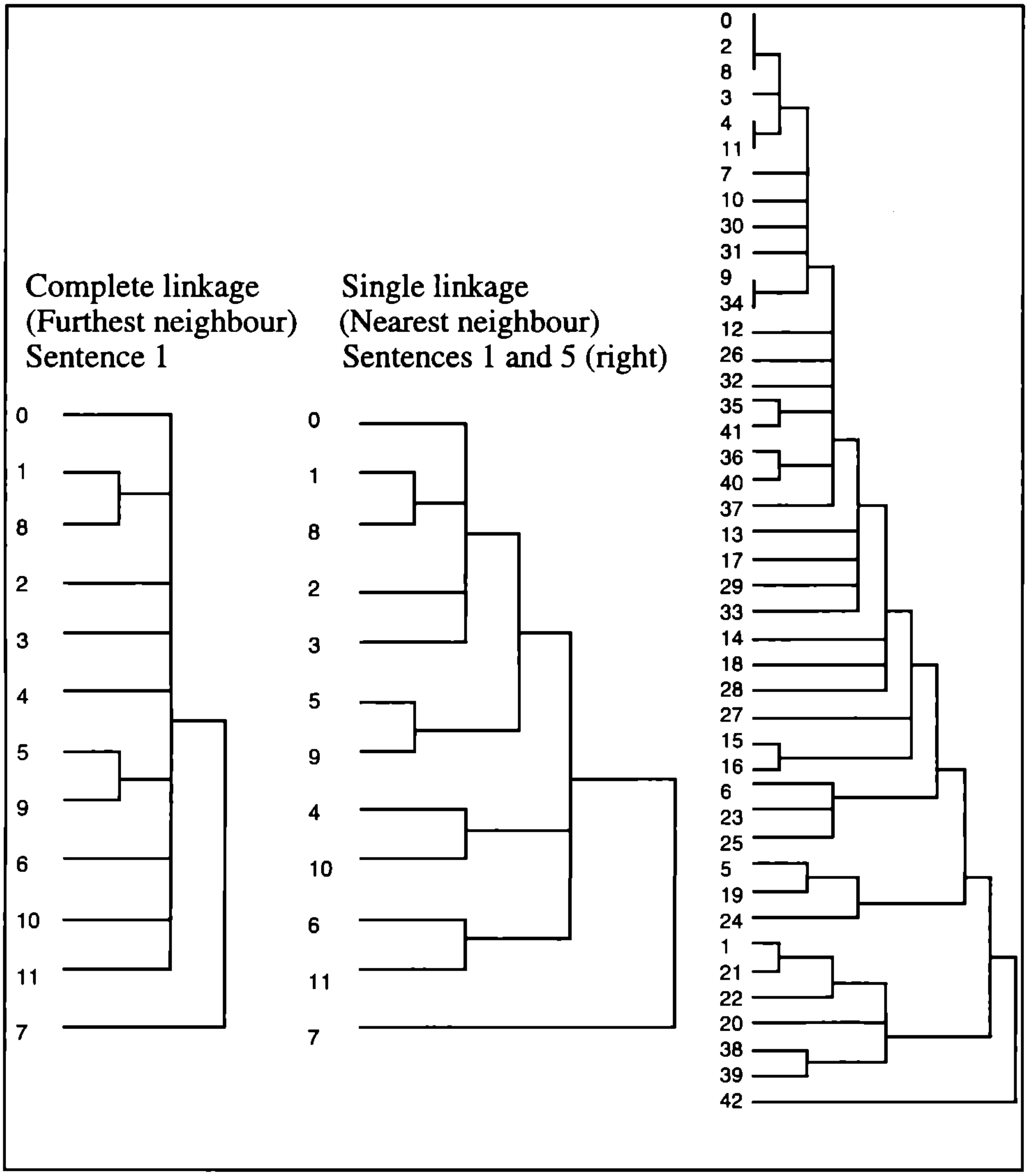
TABLE 4: Translation frequencies for 5: *"Show me the names of airlines in Atlanta"*.

0	10	visa mig namnen på flygbolagen i atlanta
1	6	vilka flygbolag finns i atlanta
2	6	visa mig namnen på flygbolag i atlanta
3	3	visa mig namnen på alla flygbolag i atlanta
4	3	visa namnen på flygbolagen i atlanta
5	2	vilka flygbolag flyger på atlanta
6	2	visa alla flygbolag i atlanta
7	2	visa mig flygbolagen i atlanta
8	2	visa mig namn på flygbolag i atlanta
9	2	visa mig namnen på flyglinjer i atlanta
10	2	visa namnen på alla flygbolag i atlanta
11	2	visa namnen på flygbolag i atlanta
12	1	ge mig namnen på flygbolag representerade i atlanta
13	1	ge mig namnen på flyglinjerna i atlanta
14	1	jag vill veta namnen på flygbolag i atlanta
15	1	kan jag få namnen på flygbolag i atlanta
16	1	kan jag få se namn på flygbolag i atlanta
17	1	vad är namnen på flygbolagen i atlanta
18	1	var god visa namnen på flyglinjerna i atlanta
19	1	vilka bolag flyger på atlanta
20	1	vilka bolag har kontor i atlanta
21	1	vilka flygbolag finns representerade i atlanta
22	1	vilka flygbolag trafikerar atlanta
23	1	visa alla bolag representerade i atlanta
24	1	visa alla flygbolag som flyger på atlanta
25	1	visa mig all flygrutter i atlanta
26	1	visa mig alla namn av flygbolag i atlanta
27	1	visa mig bolagen som flyger på atlanta
28	1	visa mig name på alla flyglinjer i atlanta
29	1	visa mig namnen av luftlinjer i atlanta
30	1	visa mig namnen för flygbolag i atlanta
31	1	visa mig namnen på bolagen i atlanta
32	1	visa mig namnen på de flygbolag som finns i atlanta
33	1	visa mig namnen på flyglinjer till atlanta
34	1	visa mig namnen på flyglinjerna i atlanta
35	1	visa mig namnen på flygrutterna i atlanta
36	1	visa mig namnen på linjer i atlanta
37	1	visa mig namnet på alla flygbolag i atlanta
38	1	visa mig vilka bolag som finns i atlanta
39	1	visa mig vilka flygbolag som finns i atlanta
40	1	visa namn på linjer i atlanta
41	1	visa namnen på flygrutter i atlanta
42	1	visa upp flyglinjer som avgår från atlanta

TABLE 5: Bigram probabilities for sentence 12

0,60	jag behöver en flight från Pittsburgh nästa måndag som är framme i Fort Worth före klockan tio
0,60	jag behöver en flight från Pittsburgh nästa måndag som anländer i Fort Worth före tio på förmiddagen
0,58	jag vill ha en flight från Pittsburgh nästa måndag som anländer i Fort Worth före klockan tio
0,58	jag behöver flyga från Pittsburgh nästa måndag och komma fram till Fort Worth före tio på morgonen
0,56	jag vill ha ett flyg från Pittsburgh nästa måndag som är framme i Fort Worth före klockan tio på morgonen
	...
0,27	jag behöver en biljett från Pittsburgh framme i Fort Worth innan tio nästa måndag
0,26	jag söker en flight till Pittsburgh nästkommande måndag som beräknas vara framme före klockan tio på morgonen
0,19	nästa måndag behöver jag flyga från Pittsburgh till Fort Worth så att jag anländer före klockan tio

FIGURE 4: Dendrograms for two different clustering methods, sentences 1 and 5.


Complete linkage
(Furthest neighbour)
Sentence 1
0
1
8
2
3
4
5
9
6
10
11
7
Single linkage
(Nearest neighbour)
Sentences 1 and 5 (right)
0
1
8
2
3
5
9
4
10
6
11
7
0
2
8
3
4
11
7
10
30
31
9
34
12
26
32
35
41
36
40
37
13
17
29
33
14
18
28
27
15
16
6
23
25
5
19
24
1
21
22
20
38
39
42