Flow Network Models for Word Alignment and Terminology Extraction from Bilingual Corpora

Éric Gaussier
Xerox Research Centre Europe 6, Chemin de Maupertuis 38240 Meylan F.
Eric.Gaussier@xrce.xerox.com

Abstract

This paper presents a new model for word alignments between parallel sentences, which allows one to accurately estimate different parameters, in a computationally efficient way. An application of this model to bilingual terminology extraction, where terms are identified in one language and guessed, through the alignment process, in the other one, is also described. An experiment conducted on a small English-French parallel corpus gave results with high precision, demonstrating the validity of the model.

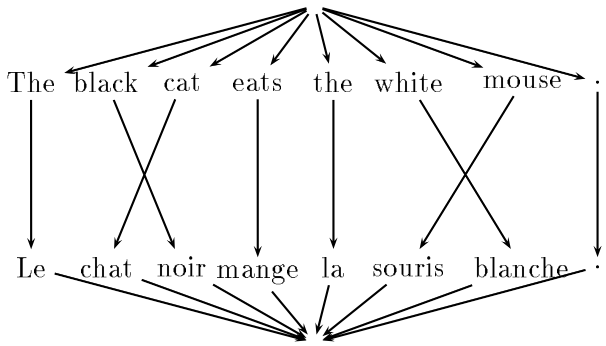
1 Introduction

Early works, (Gale and Church, 1993; Brown et al., 1993), and to a certain extent (Kay and Rőscheisen, 1993), presented methods to extract bilingual lexicons of words from a parallel corpus, relying on the distribution of the words in the set of parallel sentences (or other units). (Brown et al., 1993) then extended their method and established a sound probabilistic model series, relying on different parameters describing how words within parallel sentences are aligned to each other. On the other hand, (Dagan et al., 1993) proposed an algorithm, borrowed to the field of dynamic programming and based on the output of their previous work, to find the best alignment, subject to certain constraints, between words in parallel sentences. A similar algorithm was used by (Vogel et al., 1996). Investigating alignments at the sentence level allows to clean and to refine the lexicons otherwise extracted from a parallel corpus as a whole, pruning what (Melamed, 1996) calls "indirect associations".

Now, what differentiates the models and algorithms proposed are the sets of parameters and constraints they rely on, their ability to find an appropriate solution under the constraints defined and their ability to nicely integrate new parameters. We want to present here a model of the possible alignments in the form of flow networks. This representation allows to define different kinds of alignments and to find the most probable or an approximation of this most probable alignment, under certain constraints. Our procedure presents the advantage of an accurate modelling of the possible alignments, and can be used on small corpora. We will introduce this model in the next section. Section 3 describes a particular use of this model to find term translations, and presents the results we obtained for this task on a small corpus. Finally, the main features of our work and the research directions we envisage are summarized in the conlcusion.

2 Alignments and flow networks

Let us first consider the following aligned sentences, with the actual alignment beween words[1]:



Assuming that we have probabilities of associating English and French words, one way to find the preceding alignment is to search for the most

[1] All the examples consider English and French as the source and target languages, even though the method we propose is independent of the language pair under consideration

probable alignment under the constraints that any given English (resp. French) word is associated to one and only one French (resp. English) word. We can view a connection between an English and a French word as a flow going from an English to a French word. The preceding constraints state that the outgoing flow of an English word and the ingoing one of a French word must equal 1. We also have connections entering the English words, from a source, and leaving the French ones, to a sink, to control the flow quantity we want to go through the words.

2.1 Flow networks

We meet here the notion of flow networks that we can formalise in the following way (we assume that the reader has basic notions of graph theory).

Definition 1: let $G = (V, E)$ be a directed connected graph with m edges. A **flow** in G is a vector

$$\varphi = (\varphi_1, \varphi_2, ..., \varphi_m)^T \in R^m$$

(where T denotes the transpose of a matrix) such as, for each vertex $i \in V$:

$$\sum_{u \in \omega^+(i)} \varphi_u = \sum_{u \in \omega^-(i)} \varphi_u \tag{1}$$

where $\omega^+(i)$ denotes the set of edges entering vertex i, whereas $\omega^-(i)$ is the set of edges leaving vertex i.

We can, furthermore, associate to each edge u of $G = (V, E)$ two numbers, b_u and c_u with $b_u \leq c_u$, which will be called the lower capacity bound and the upper capacity bound of the edge.

Definition 2: let $G = (V, E)$ be a directed connected graph with lower and upper capacity bounds. We will say that a flow φ in G is a **feasible flow** in G if it satisfies the following capacity constraints:

$$\forall u \in E, b_u \leq \varphi_u \leq c_u \tag{2}$$

Finally, let us associate to each edge u of a directed connected graph $G = (V, E)$ with capacity intervals $[b_u; c_u]$ a cost γ_u, representing the cost (or inversely the probability) to use this edge in a flow. We can define the total cost, $\gamma \times \varphi$, associated to a flow φ in G as follows:

$$\gamma \times \varphi = \sum_{u \in E} \gamma_u \times \varphi_u \tag{3}$$

Definition 3: let $G = (V, E)$ be a connected graph with capacity intervals $[b_u; c_u]$, $u \in E$ and costs γ_u, $u \in E$. We will call **minimal cost flow** the feasible flow in G for which $\gamma \times \varphi$ is minimal.

Several algorithms have been proposed to compute the minimal cost flow when it exists. We will not detail them here but refer the interested reader to (Ford and Fulkerson, 1962; Klein, 1967).

2.2 Alignment models

Flows and networks define a general framework in which it is possible to model alignments between words, and to find, under certain constraints, the best alignment. We present now an instance of such a model, where the only parameters involved are **association probabilities** between English and French words, and in which we impose that any English, respectively French word, has to be aligned with one and only one French, resp. English, word, possibly empty. We can, of course, consider different constraints. The constraints we define, though they would yield to a complex computation for the EM algorithm, do not privilege any direction in an underlying translation process.

This model defines for each pair of aligned sentences a graph $G(V, E)$ as follows:

- V comprises a source, a sink, all the English and French words, an empty English word, and an empty French word,
- E comprises edges from the source to all the English words (including the empty one), edges from all the French words (including the empty one) to the sink, an edge from the sink to the source, and edges from all English words (including the empty one) to all the French words (including the empty one)[2].
- from the source to all possible English words (excluding the empty one), the capacity interval is [1;1],

[2] The empty words account for the fact that words may not be aligned with other ones, i.e. they are not explicitely translated for example.

- from the source to the empty English word, the capacity interval is $[0; max(l_e, l_f)]$, where l_f is the number of French words, and l_e the number of English ones,
- from the English words (including the empty one) to the French words (including the empty one), the capacity interval is [0;1],
- from the French words (excluding the empty one) to the sink, the capacity interval is [1;1],
- from the empty French word to the sink, the capacity interval is $[0; max(l_e, l_f)]$,
- from the sink to the source, the capacity interval is $[0; max(l_e, l_f)]$.

Once such a graph has been defined, we have to assign cost values to its edges, to reflect the different association probabilities. We will now see how to define the costs so as to relate the minimal cost flow to a best alignment. Let a be an alignment, under the above constraints, between the English sentence e_s, and the French sentence f_s. Such an alignment a can be seen as a particular relation from the set of English words with their positions, including empty words, to the set of French words with their positions, including empty words (in our framework, it is formally equivalent to consider a single empty word with larger upper capacity bound or several ones with smaller upper capacity bounds; for the sake of simplicity in the formulas, we consider here that we add as many empty words as necessary in the sentences to end up with two sentences containing $l_e + l_f$ words). An alignment thus connects each English word, located in position i, e_i, to a French word, in position j, f_j. We consider that the probability of such a connection depends on two distinct and independent probabilities, the one of linking two positions, $p(a_p(i) = a_i)$, and the one of linking two words, $p(a_w(e_i) = f_{a_i})$. We can then write:

$$P(a, e_s, f_s) = \prod_{i=1}^{l_e+l_f} p(a_p(i) = a_i | (a, e, f)_1^{i-1}) \prod_{i=1}^{l_e+l_f} p(a_w(e_i) = f_{a_i} | (a, e, f)_1^{i-1}) \quad (4)$$

where $P(a, e_s, f_s)$ is the probability of observing the alignment a together with the English and French sentences, e_s and f_s, and $(a, e, f)_1^{i-1}$ is a shorthand for $(a_1, .., a_{i-1}, e_1, .., e_{i-1}, f_{a_1}, .., f_{a_{i-1}})$.

Since we simply rely in this model on association probabilities, that we assume to be independent, the only dependencies lying in the possibilities to associate words across languages, we can simplify the above formula and write:

$$P(a, e_s, f_s) = \prod_{i=1}^{l_e+l_f} p(e_i, f_{a_i} | a_1^{i-1}) \quad (5)$$

where a_1^{i-1} is a shorthand for $(a_1, .., a_{i-1})$. $p(e_i, f_{a_i})$ is a shorthand for $p(a_w(e_i) = f_{a_i})$ that we will use throughout the article. Due to the constraints defined, we have: $p(e_i, f_{a_i} | a_1^i) = 0$ if $a_i \in a_1^{i-1}$, and $p(e_i, f_{a_i})$ otherwise.

Equation (5) shows that if we define the cost associated to each edge from an English word e_i (excluding the empty word) to a French word f_j (excluding the empty word) to be $\gamma_u = -\ln p(e_i, f_j)$, the cost of an edge involving an empty word to be ϵ, an arbitrary small positive value, and the cost of all the other edges (i.e. the edges from SoP and SiP) to be 1 for example, then the minimal cost flow defines the alignment a for which $P(a, e_s, f_s)$ is maximum, under the above constraints and approximations.

We can use the following general algorithm based on maximum likelihood under the maximum approximation, to estimate the parameters of our model:

1. set some initial value to the different parameters of the model,
2. for each sentence pair in the corpus, compute the best alignment (or an approximation of this alignment) between words, with respect to the model, and update the counts of the different parameters with respect to this alignment (the maximum likelihood estimators for model free distributions are based on relative frequencies, conditioned by the set of best alignments in our case),
3. go back to step 2 till an end condition is reached.

This algorithm converges after a few iterations. Here, we have to be carefull with step 1. In particular, if we consider at the beginning of the process all the possible alignments to be equiprobable, then all the feasible flows are minimal cost flows. To avoid this situation, we have to start with initial probabilities which make use of the fact that some associations, occurring more often in the corpus, should have a larger probability. Probabilities based on relative frequencies, or derived from the measure defined in (Dunning, 1993), for example, allow to take this fact into account.

We can envisage more complex models, including distortion parameters, multiword notions, or information on part-of-speech, information derived from bilingual dictionaries or from thesauri. The integration of new parameters is in general straigthforward. For multiword notions, we have to replace the capacity values of edges connected to the source and the sink with capacity intervals, which raises several issues that we will not address in this paper. We rather want to present now an application of the flow network model to multilingual terminology extraction.

3 Multilingual terminology extraction

Several works describe methods to extract terms, or candidate terms, in English and/or French (Justeson and Katz, 1995; Daille, 1994; Nkwenti-Azeh, 1992). Some more specific works describe methods to align noun phrases within parallel corpora (Kupiec, 1993). The underlying assumption beyond these works is that the monolingually extracted units correspond to each other cross-lingually. Unfortunately, this is not always the case, and the above methodology suffers from the weaknesses pointed out by (Wu, 1997) concerning parse-parse-match procedures.

It is not however possible to fully reject the notion of grammar for term extraction, in so far as terms are highly characterized by their internal syntactic structure. We can also admit that lexical affinities between the diverse constituents of a unit can provide a good clue for termhood, but lexical affinities, or otherwise called collocations, affect different linguistic units that need anyway be distinguished (Smadja, 1992).

Moreover, a study presented in (Gaussier, 1995) shows that terminology extraction in English and in French is not symmetric. In many cases, it is possible to obtain a better approximation for English terms than it is for French terms. This is partly due to the fact that English relies on a composition of Germanic type, as defined in (Chuquet and Paillard, 1989) for example, to produce compounds, and of Romance type to produce free NPs, whereas French relies on Romance type for both, with the classic PP attachment problems.

These remarks lead us to advocate a mixed model, where candidate terms are identified in English and where their French correspondent is searched for. But since terms constitute rigid units, lying somewhere between single word notions and complete noun phrases, we should not consider all possible French units, but only the ones made of consecutive words.

3.1 Model

It is possible to use flow network models to capture relations between English and French terms. But since we want to discover French units, we have to add extra vertices and nodes to our previous model, in order to account for all possible combinations of consecutive French words. We do that by adding several layers of vertices, the lowest layer being associated with the French words themselves, and each vertex in any upper layer being linked to two consecutive vertices of the layer below. The uppest layer contains only one vertex and can be seen as representing the whole French sentence. We will call a **fertility graph** the graph thus obtained. Figure 1 gives an example of part of a fertility graph (we have shown the flow values on each edge for clarity reasons; the brackets delimit a nultiword candidate term; we have not drawn the whole fertility graph encompassing the French sentence, but only part of it, the one encompassing the unit *largeur de bande utilisée*, where the possible combinations of consecutive words are represented by A, B, and C). Note that we restrict ourselves to lexical words (nouns, verbs, adjectives and adverbs), not trying to align grammatical words. Furthermore, we rely on lemmas rather than inflected froms, thus enabling us to conflate in one form all the variants of a verb for example (we have keeped

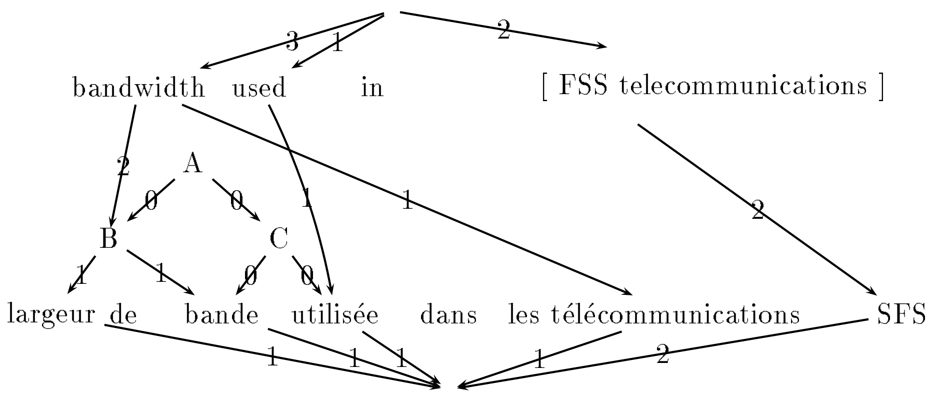


Figure 1: Pseudo-alignment within a fertility graph

inflected forms in our figures for readability reasons).

The minimal cost flow in the graphs thus defined may not be directly usable. This is due to two problems:

1. first, we can have ambiguous associations: in figure 1, for example, the association between *bandwidth* and *largeur de bande* can be obtained through the edge linking these two units (type 1), or through two edges, one from *bandwidth* to *largeur de bande*, and one from *bandwidth* to either *largeur* or *bande* (type 2), or even through the two edges from *bandwidth* to *largeur* and *bande* (type 3),

2. secondly, there may be conflicts between connections: in figure 1 both *largeur de bande* and *télécommunications* are linked to *bandwidth* even though they are not contiguous.

To solve ambiguous associations, we simply replace each association of type 2 or 3 by the equivalent type 1 association[3]. For conflicts, we use the following heuristics: first select the conflicting edge with the lowest cost and assume that the association thus defined actually occurred, then rerun the minimal cost flow algorithm with this selected edge fixed once and for all, and redo these two steps until there is no more conflicting edges, replacing type 2 or 3 associations as above each time it is necessary.

Finally, the alignment obtained in this way will be called a **solved alignment**[4].

3.2 Experiment

In order to test the previous model, we selected a small bilingual corpus consisting of 1000 aligned sentences, from a corpus on satellite telecommunications. We then ran the following algorithm, based on the previous model:

1. tag and lemmatise the English and French texts, mark all the English candidate terms using morpho-syntactic rules encoded in regular expressions,

2. build a first set of association probabilities, using the likelihood ratio test defined in (Gaussier, 1995),

3. for each pair of aligned sentences, construct the fertility graph allowing a candidate term of length n to be aligned with units of lenth (n-2) to (n+2), define the

[3] We can formally define an equivalence relation, in terms of the associations obtained, but this is beyond the scope of this paper.

[4] Once the solved alignment is computed, it is possible to determine the word associations between aligned units, through the application of the process described in the previous section with multiword notions.

costs of edges linking English vertices to French ones as the opposite of the logarithm of the normalised sum of probabilities of all possible word associations defined by the edge (for the edge between *multiple (e1) access (e2)* to the French unit *accès (f1) mulitple (f2)* it is $\frac{1}{4}\left(\sum_{i,j} p(ei, fj)\right)$), all the other edges receive an arbitrary cost value, compute the solved alignment, and increment the count of the associations obtained by overall value of the solved alignment,

4. select the fisrt 100 unit associations according to their count, and consider them as valid. Go back to step 2, excluding from the search space the associations selected, till all associations have been extracted.

3.3 Results

To evaluate the results of the above procedure, we manually checked each set of associations obtained after each iteration of the process, going from the first 100 to the first 500 associations. We considered an association as being correct if the French expression is a proper translation of the English expression. The following table gives the precision of the associations obtained.

N. Assoc.	Prec.
100	98
200	97
300	96
400	95
500	90

Table 1: General results

The associations we are faced with represent different linguistic units. Some consist of single content words, whereas others represent multiword expressions. One of the particularity of our process is precisely to automatically identify multiword expressions in one language, knowing units in the other one. With respect to this task, we extracted the first two hundred multiword expressions from the associations above, and then checked wether they were valid or not. We obtained the following results:

N. Assoc.	Prec.
100	97
200	94

Table 2: Multiword notion results

As a comparison, (Kupiec, 1993) obtained a precision of 90% for the first hundred associations between English and French noun phrases, using the EM algorithm. Our experiments with a similar method showed a precision around 92% for the first hundred associations on a set of aligned sentences comprising the one used for the above experiment.

An evaluation on single words, showed a precision of 98% for the first hundred and 97% for the first two hundred. But these figures should be seen in fact as lower bounds of actual values we can get, in so far as we have not tried to extract single word associations from multiword ones. Here is an example of associations obtained.

telecommunication satellite
satelllite de télécommunication
communication satellite
satelllite de télécommunication
new satellite system
nouveau système de satellite
système de satellite nouveau
système de satellite entièrement nouveau
operating fss telecommunication link
exploiter la liason de télécommunication du sfs
implement *mise en oeuvre*
wavelength *longueur d'onde*
offer *offrir, proposer*
operation *exploitation, opération*

The empty words (prepositions, determiners) were extracted from the sentences. In all the cases above, the use of prepositions and determiners was consistent all over the corpus. There are cases where two French units differ on a preposition. In such a case, we consider that we have two possible different translations for the English term.

4 Conclusion

We presented a new model for word alignment based on flow networks. This model allows us to integrate different types of constraints in the search for the best word alignment within aligned sentences. We showed how this model can be applied to terminology extraction, where candidate terms are extracted in one language, and discovered, through the alignment process, in the other one. Our procedure presents three main differences over other approaches: we do not force term translations to fit within specific patterns, we consider the whole sentences, thus enabling us to remove some ambiguities, and we rely on the association probabilities of the units as a whole, but also on the association probabilities of the elements within these units.

The main application of the work we have described concerns the extraction of bilingual lexicons. Such extracted lexicons can be used in different contexts: as a source to help lexicographers build bilingual dictionaries in technical domains, or as a resource for machine aided human translation systems. In this last case, we can envisage several ways to extend the notion of translation unit in translation memory systems, as the one proposed in (Langé et al., 1997).

5 Acknowledgements

Most of this work was done at the IBM-France Scientific Centre during my PhD research, under the direction of Jean-Marc Langé, to whom I express my gratitude. Many thanks also to Jean-Pierre Chanod, Andeas Eisele, David Hull, and Christian Jacquemin for useful comments on earlier versions.

References

Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer. 1993. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

H. Chuquet and M. Paillard. 1989. *Approche linguistique des problèmes de traduction anglais-français*. Ophrys.

Ido Dagan, Kenneth W. Church, and William A. Gale. 1993. Robust bilingual word alignment for machine aided translation. In *Proceedings of the Workshop on Very Large Corpora*.

Béatrice Daille. 1994. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Ph.D. thesis, Univ. Paris 7.

T. Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1).

L.R. Ford and D.R. Fulkerson. 1962. *Flows in networks*. Princeton University Press.

William Gale and Kenneth Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1).

Éric Gaussier. 1995. *Modèles statistiques et patrons morphosyntaxiques pour l'extraction de lexiques bilingues de termes*. Ph.D. thesis, Univ. Paris 7.

John S. Justeson and Slava M. Katz. 1995. Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1).

Martin Kay and M. Rőscheisen. 1993. Text-translation alignment. *Computational Linguistics*, 19(1).

M. Klein. 1967. A primal method for minimal cost flows, with applications to the assignment and transportation problems. *Management Science*.

Julian Kupiec. 1993. An algorithm for finding noun phrase correspondences in bilingual corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*.

Jean-Marc Langé, Éric Gaussier, and Béatrice Daill. 1997. Bricks and skeletons: some ideas for the near future of maht. *Machine Translation*, 12(1).

Dan I. Melamed. 1996. Automatic construction of clean broad-coverage translation lexicons. In *Proceedings of the Second Conference of the Association for Machine Translation in the Americas (AMTA)*.

Basile Nkwenti-Azeh. 1992. Positional and combinational characteristics of satellite communications terms. Technical report, CCl-UMIST, Manchester.

Frank Smadja. 1992. How to compile a bilingual collocational lexicon automatically. In *Proceedings of AAAI-92 Workshop on Statistically-Based NLP techniques*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. Hmm-based word alignment in statistical translation. In *Proceedings of the Sixteenth International Conference on Computational Linguistics*.

Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational Linguistics*, 23(3).