Extracting Parallel Fragments from Comparable Corpora for Data-to-text Generation

Anja Belz **Eric Kow**
Natural Language Technology Group
School of Computing, Mathematical and Information Sciences
University of Brighton
Brighton BN2 4GJ, UK
{asb,eykk10}@bton.ac.uk

Abstract

Building NLG systems, in particular statistical ones, requires parallel data (paired inputs and outputs) which do not generally occur naturally. In this paper, we investigate the idea of automatically extracting parallel resources for data-to-text generation from *comparable* corpora obtained from the Web. We describe our comparable corpus of data and texts relating to British hills and the techniques for extracting paired input/output fragments we have developed so far.

1 Introduction

Starting with Knight, Langkilde and Hatzivassiloglou's work on Nitrogen and its successor Halogen (Knight and Hatzivassiloglou, 1995; Knight and Langkilde, 2000), NLG has over the past 15 years moved towards using statistical techniques, in particular in surface realisation (Langkilde, 2002; White, 2004), referring expression generation (most of the sytems submitted to the TUNA and GREC shared task evaluation challenges are statistical, see Gatt et al. (2008), for example), and data-to-text generation (Belz, 2008).

The impetus for introducing statistical techniques in NLG can be said to have originally come from machine translation (MT),[1] but unlike MT, where parallel corpora of inputs (source language texts) and outputs (translated texts) occur naturally at least in some domains,[2] NLG on the whole has to use manually created input/output pairs.

Data-to-text generation (D2T) is the type of NLG that perhaps comes closest to having naturally occuring inputs and outputs at its disposal. Work in D2T has involved different domains including generating weather forecasts from meteorological data (Sripada et al., 2003), nursing reports from intensive care data (Portet et al., 2009), and museum exhibit descriptions from database records (Isard et al., 2003; Stock et al., 2007); types of data include dynamic time-series data (e.g. medical data) and static database entries (museum exhibits).

While data and texts in the three example domains cited above do occur naturally, two factors mean they cannot be used directly as example corpora or training data for building D2T systems: one, most are not freely available to researchers (e.g. by simply being available on the Web), and two, more problematically, for the most part, there is no direct correspondence between inputs and outputs as there is, say, between a source language text and its translation. On the whole, naturally occurring resources of data and related texts are not strictly parallel, but are merely what has become known as *comparable* in the MT literature, with only a subset of data having corresponding text fragments, and other text fragments having no obvious corresponding data items. Moreover, data transformations may be necessary before corresponding text fragments can be identified.

In this report, we look at the possibility of automatically extracting parallel data-text fragments from comparable corpora in the case of D2T from static database records. Such a parallel data-text resource could then be used to train an existing D2T generation system, or even build a new statistical generator from scratch, e.g. using techniques from statistical MT (Belz and Kow, 2009). The steps involved in going from comparable data and text resources to generators that produce texts similar to those in the text resource are then as follows: (1) identify sources on the Web for comparable data and texts; (2) pair up data records and texts; (3) extract parallel fragments (sets of data fields paired with word strings); (4) train a D2T generator using the parallel fragments; and (5) feed data inputs to the generator which then

[1] Nitrogen was conceived as an MT system component.

[2] Canadian and European parliamentary proceedings, etc.

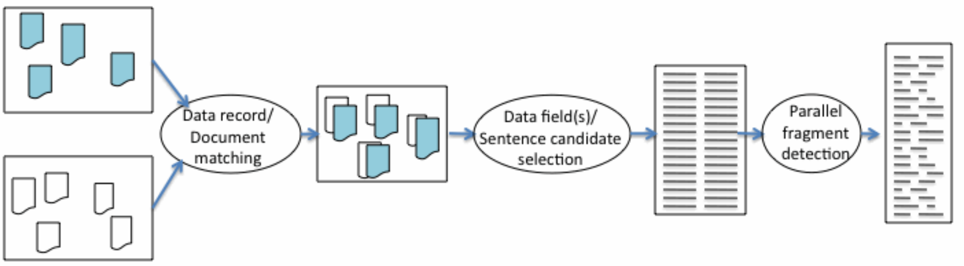

Figure 1: Overview of processing steps.

generates new texts describing them. Figure 1 illustrates steps 1–3 which this paper focuses on. In Section 3 we look at steps 1 and 2; in Section 4 at step 3. First we briefly survey related work in MT.

2 Related work in MT

In statistical MT, the expense of manually creating new parallel MT corpora, and the need for very large amounts of parallel training data, has led to a sizeable research effort to develop methods for automatically constructing parallel resources. This work typically starts by identifying comparable corpora. Much of it has focused on identifying word translations in comparable corpora, e.g. Rapp's approach was based on the simple and elegant assumption that if words A_f and B_f have a higher than chance co-occurrence frequency in one language, then two appropriate translations A_e and B_e in another language will also have a higher than chance co-occurrence frequency (Rapp, 1995; Rapp, 1999). At the other end of the spectrum, Resnik & Smith (2003) search the Web to detect web pages that are translations of each other. Other approaches aim to identify pairs of sentences (Munteanu and Marcu, 2005) or sub-sentential fragments (Munteanu and Marcu, 2006) that are parallel within comparable corpora.

The latter approach is particularly relevant to our work. They start by translating each document in the source language (SL) word for word into the target language (TL). The result is given to an information retrieval (IR) system as a query, and the top 20 results are retained and paired with the given SL document. They then obtain all sentence pairs from each pair of SL and TL documents, and discard those sentence pairs with few words that are translations of each other. To the remaining sentences they then apply a fragment detection method which tries to distinguish between source fragments that have a translation on the target side, and fragments that do not.

The biggest difference between the MT situation and the D2T situation is that in the latter sentence-aligned parallel resources exist and can be used as a starting point. E.g. Munteanu & Marcu use an existing parallel Romanian-English corpus to (automatically) create a lexicon from which is then used in various ways in their method.

In D2T we have no analogous resources to help us get started, and the methods described in this paper use no such prior knowledge.

3 A Comparable Corpus of British Hills

As a source of data, we use the Database of British Hills (BHDB) created by Chris Crocker,[3] version 11.3, which currently contains measurements and other information about 5,614 British hills. Additionally, we perform reverse geocoding via the Google Map API[4] which allows us to convert latitude and longitude information from the hills database into country and region names. We add the latter to each database entry.

On the text side, we use Wikipedia texts in the WikiProject British and Irish Hills (retrieved on 2009-11-09). There are currently 899 pages covered by this WikiProject, 242 of which are of quality category B or above.[5]

Matching up data records and documents: Matching up the data records in the BHDB with articles in Wikipedia is not trivial: not all BHDB entries have corresponding Wikipedia articles, different hills often share the same name, and the same hill can have different names and spellings.

We perform a search of Wikipedia with the hill's name as the search term, using the Mediawiki API, and then retain the top n search results returned (currently $n = 1$). The top search result is not always a correct match for the database record. We

[3] http://www.biber.fsnet.co.uk

[4] `http://code.google.com/apis/maps/`

[5] B = The article is mostly complete and without major issues, but requires some further work.

```
{ "id": 1679, "main-name-info": {"name": "Hill of Stake", "notes": "",
                                 "parent": "", "parent-notes": ""},
"alt-name-info": [], "raw-name": "Hill of Stake", "rhb-section": "27A", "area": "Ayr to River Clyde",
"height-metres": 522, "height-feet": 1713, "map-1to50k": "63", "map-1to25k": "341N", "gridref": "NS273630",
"col-gridref": "NS320527", "col-height": 33, "drop": 489, "gridref10": "NS 27360 62998", "feature": "trig point",
"observations": "", "survey": "", "date-climbed": "", "classification": "Ma,CoH,CoU",
"county-name": "Renfrewshire(CoH); Renfrewshire(CoU)", "revision": "28-Oct-2001", "comments": "",
"streetmap": "http://www.streetmap.co.uk/newmap.srf?x=227356&y=663005&z=3&sv=227356,663005&st=4&tl=˜&bi=˜&lu=N&ar=n",
"ordanancesurvey-map": "http://getamap.ordnancesurvey.co.uk/getamap/frames.htm?mapAction=gaz&gazName=g&gazString=NS273630",
"x-coord": 227356, "y-coord": 663005, "latitude": 55.82931,
"longitude": -4.75789, "country": "Scotland", "region": "Renfrewshire" }
```

Hill of Stake is a hill on the boundary between North Ayrshire and Renfrewshire , Scotland . It is 522 metres (1712 feet) high . It is one of the Marilyns of Lowland Scotland . It is the highest point of the relatively low-lying county of Renfrewshire and indeed the entire Clyde Muirshiel Regional Park of which it is a part .

Table 1: Output of step 1: data record from British Hills DB and matched Wikipedia text (Hill of Stake).

manually selected the pairs we are confident are a correct match. This left us with 759 matched pairs out of a possible 899.

Table 1 shows an example of an automatically matched database entry and Wikipedia article. It illustrates the non-parallelism discussed in the preceding section; e.g. there is no information in the database corresponding to the last sentence.

4 Towards a Parallelised Corpus

4.1 Aligning data fields and sentences

In the second processing step, we pair up data fields and sentences. Related methods in MT have translation lexicons and thesauri that can be used as bridges between SL and TL texts, but there is no equivalents in NLG. Our current method associates each data field with a hand-written 'match predicate'. For example, the match predicate for `height-metres` returns True if the sentence contains the words 'X metres' (among other patterns), where X is some number within 5% of the height of the hill in the database. We retain only the sentences that match at least one data field. Table 2 shows what the data field/sentence alignment procedure outputs for the Hill of Stake.

4.2 Identifying Parallel Fragments

While it was fine for step 2 to produce some rough matches, in step 3, parallel fragment detection, the aim is to retain only those parts of a sentence that can be said to realise some data field(s) in the set of data fields with which it has been matched.

Computing data-text associations: Following some preprocessing of sentences where each occurrence of a hill's name and height is replaced by lexical class tokens _NAME_, _HEIGHT_METRES_ or _HEIGHT_FEET_, the first step is to construct a kind of lexicon of pairs (d, w) of data fields d and words w, such that w is often seen in the realisation of d. For this purpose we adapt Munteanu & Marcu's (2006) method for (language to language) lexicon construction. For this purpose we compute a measure of the strength of association between data fields and words; we use the G^2 log-likelihood ratio which has been widely used for this sort of purpose (especially lexical association) since it was introduced to NLP (Dunning, 1993). Following Moore (2004a) rather than Munteanu & Marcu, our current notion of cooccurrence is that a data field and word cooccur if they are present in the same pair of data fields and sentence (as identified by the method described in Section 4.1 above). We then obtain counts for the number of times each word cooccurs with each data field, and the number of times it occurs without the data field being present (and conversely). This allows us to compute the G^2 score, for which we use the formulation from Moore (2004b) shown in Figure 2.

If the G^2 score for a given (d, w) pair is greater than $p(d)p(w)$, then the association is taken to be positive, i.e. w is likely to be a realisation of d, otherwise the association is taken to be negative, i.e. w is likely not to be part of a realisation of d.

For each d we then convert G^2 scores to probabilities by dividing G^2 by the appropriate normalising factor (the sum over all negative G^2 scores for d for obtaining the negative association probabilities, and analogously for positive associations). Table 3 shows the three words with the highest positive association probabilities for each of our six data fields. Note that these are not the three most likely alternative 'translations' of each data key, but rather the three words which are most likely to be part of a realisation of a data field, if seen in conjunction with it.

`"main-name-only": "Hill of Stake",`
`"country": "Scotland"`
	NAME is a hill on the boundary between North Ayrshire and Renfrewshire, Scotland.

`"height-metres": 522,`
`"height-feet": 1713`
	It is _HEIGHT_METERS_ metres (_HEIGHT_FEET_ feet) high.

`"country": "Scotland",`
`"classification": ["Ma", "CoH", "CoU"]`
	It is one of the Marilyns of Lowland Scotland.

`"main-name-only": "Hill of Stake"`	It is the highest point of the relatively low-lying county of Renfrewshire and indeed the entire Clyde Muirshiel Regional Park of which it is a part.

Table 2: Output of step 2: aligned data fields and sentences, for Hill of Stake.

$$2N\left(p(d,w)log\frac{p(d,w)}{p(d)p(w)}+p(d,\neg w)log\frac{p(d,\neg w)}{p(d)p(\neg w)}+p(\neg d,w)log\frac{p(\neg d,w)}{p(\neg d)p(w)}+p(\neg d,\neg w)log\frac{p(\neg d,\neg w)}{p(\neg d)p(\neg w)}\right)$$

Figure 2: Formula for computing G^2 from Moore (2004b) (N is the sample size).

Data key d	Word w	$P^+(w\|d)$
`main-name-only`	_NAME_ a in	0.1355 0.0742 0.0660
`classification`	as adjoining qualifies	0.0412 0.0193 0.0177
`region`	District Lake area	0.1855 0.1661 0.1095
`country`	in _NAME_ Scotland	0.1640 0.1122 0.0732
`height-metres`	metres m height	0.1255 0.0791 0.0679
`height-feet`	feet _HEIGHT_METERS_ (	0.1511 0.0974 0.0900

Table 3: Data keys with 3 most likely words.

Identifying realisations: The next step is to apply these probabilities to identify those parts of a sentence that are likely to be a valid realisation of the data fields in the input. In Figure 3 we plot the positive and negative association probabilities for one of the sentences from our running example, Hill of Stake. The light grey graph represents the association probabilities between each word in the sentence and `height-feet`, the dark grey line those between the words in the sentence and `height-metres`. We plot the negative association probabilities simply by multiplying each by -1.

The part of the sentence that one would want to extract as a possible realisation of { `height-metres`, `height-feet` }, namely "_HEIGHT_METRES_ metres (_HEIGHT_FEET_ feet) high", shows up clearly as a sequence of relatively strong positive association values. Our current approach identifies such contiguous positive association scores and extracts the corresponding sentence fragments. This works well in many cases, but is too simple as a general approach; we are currently developing this method further.

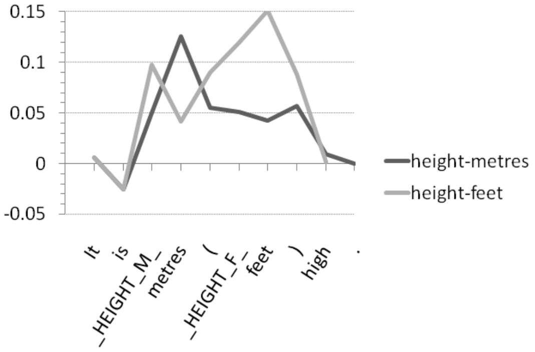

Figure 3: Positive and negative association probabilities plotted against the words they were computed for.

5 Concluding Remarks

In this paper we have been interested in the problem of automatically obtaining parallel corpora for data-to-text generation. We presented our comparable corpus of 759 paired database entries and human-authored articles about British Hills. We described the three techniques which we have implemented so far and which we combine to extract parallel data-text fragments from the corpus: (i) identification of candidate pairs of data fields and sentences; (ii) computing scores for the strength of association between data and words; and (iii) identifying sequences of words in sentences that have positive association scores with the given data fields.

References

Anja Belz and Eric Kow. 2009. System building cost vs. output quality in data-to-text generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*, pages 16–24.

A. Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Natural Language Engineering*, 14(4):431–455.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 1:61–74.

A. Gatt, A. Belz, and E. Kow. 2008. The TUNA Challenge 2008: Overview and evaluation results. In *Proceedings of the 5th International Natural Language Generation Conference (INLG'08)*, pages 198–206.

A. Isard, J. Oberlander, I. Androutsopoulos, and C. Matheson. 2003. Speaking the users' languages. 18(1):40–45.

K. Knight and V. Hatzivassiloglou. 1995. Two-level, many-paths generation. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL '95)*.

Kevin Knight and Irene Langkilde. 2000. Preserving ambiguity in generation via automata intersection. In *Proceedings of AAAI/IAAI*, pages 697–702.

I. Langkilde. 2002. An empirical verification of coverage and correctness for a general-purpose sentence generator. In *Proc. 2nd International Natural Language Generation Conference (INLG '02)*.

Robert C. Moore. 2004a. Improving ibm word-alignment model 1. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 519–526.

Robert C. Moore. 2004b. On log-likelihood-ratios and the significance of rare events. In *Proceedings of the 9th Converence on Empirical Methods in Natural Language Processing (EMNLP'04)*, pages 333–340.

Dragos Munteanu and Daniel Marcu. 2005. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31:477–504.

Dragos Stefan Munteanu and Daniel Marcu. 2006. Extracting parallel sub-sentential fragments from non-parallel corpora. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (COLING-ACL'06)*, pages 81–88, Morristown, NJ, USA. Association for Computational Linguistics.

F. Portet, E. Reiter, A. Gatt, J. Hunter, S. Sripada, Y. Freer, and C. Sykes. 2009. Automatic generation of textual summaries from neonatal intensive care data. *Artificial Intelligence*, 173:789–816.

Reinhard Rapp. 1995. Identifying word translations in non-parallel texts. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 320–322, Morristown, NJ, USA. Association for Computational Linguistics.

Reinhard Rapp. 1999. Automatic identification of word translations from unrelated english and german corpora. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 519–526, Morristown, NJ, USA. Association for Computational Linguistics.

Philip Resnik and Noah Smith. 2003. The web as a parallel corpus. *Computational Linguistics*, 29:349–380.

S. Sripada, E. Reiter, J. Hunter, and J. Yu. 2003. Exploiting a parallel text-data corpus. In *Proceedings of Corpus Linguistics 2003*, pages 734–743.

Oliviero Stock, Massimo Zancanaro, Paolo Busetta adn Charles Callaway, Anbtonio Krüger, Michael Kruppa, Tsvi Kuflik, Elena Not, and Cesare Rocchi. 2007. Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User Modeling and User-Adapted Interaction*, 17(3):257–304.

M. White. 2004. Reining in CCG chart realization. In A. Belz, R. Evans, and P. Piwek, editors, *Proceedings INLG'04*, volume 3123 of *LNAI*, pages 182–191. Springer.