Unsupervised Alignment of Comparable Data and Text Resources

Anja Belz **Eric Kow**
School of Computing, Engineering and Mathematics
University of Brighton
Brighton BN2 4GJ, UK
{A.S.Belz,E.Y.Kow}@brighton.ac.uk

Abstract

In this paper we investigate automatic data-text alignment, i.e. the task of automatically aligning data records with textual descriptions, such that data tokens are aligned with the word strings that describe them. Our methods make use of log likelihood ratios to estimate the strength of association between data tokens and text tokens. We investigate data-text alignment at the document level and at the sentence level, reporting results for several methodological variants as well as baselines. We find that log likelihood ratios provide a strong basis for predicting data-text alignment.

1 Introduction

Much of NLP system building currently uses aligned parallel resources that provide examples of the inputs to a system and the outputs it is intended to produce. In Machine Translation (MT), such resources take the form of sentence-aligned parallel corpora of source-language and target-language texts; in parsing and surface realisation, parse-annotated corpora of naturally occurring texts are used, where in parsing, the inputs are the sentences in the texts and the outputs are the parses represented by the annotations on the sentences, and in surface realisation, the roles of inputs and outputs are reversed.

In MT parallel resources exist, and in fact are produced in large quantities daily, and in some cases (e.g. multilingual parliamentary proceedings) are publicly available. Moreover, even if resources are created specifically for system building (e.g. NIST's OpenMT evaluations) the cost is offset by the fact that the resulting translation system can be expected to generalise to new domains to some extent.

While parse-annotated corpora are in the first instance created by hand, here too, parsers and surface realisers built on the basis of such corpora are expected to generalise beyond the immediate corpus domain.

In data-to-text generation, as in parsing, parallel resources do not occur naturally and have to be created manually. The associated cost is, however, incurred for every new task, as systems trained on a given parallel data-text resource cannot be expected to generalise beyond task and domain. Automatic data-text alignment methods, i.e. automatic methods for creating parallel data-text resources, would be extremely useful for system building in this situation, but no such methods currently exist.

In MT there have been recent efforts (reviewed in the following section) to automatically produce aligned parallel corpora from comparable resources where texts in two different languages are about similar topics, but are not translations of each other). Taking our inspiration from this work in MT, in this paper we investigate the feasibility of automatically creating aligned parallel data-text resources from comparable data and text resources available on the web. This task of automatic data-text alignment, previously unexplored as far as we are aware, is the task of automatically aligning data records with textual descriptions, such that data tokens are aligned with the word strings that describe them. For example, the data tokens `height_metres=250` might be aligned with the word string *with an altitude of 250*

Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 102–109,
49th Annual Meeting of the Association for Computational Linguistics,
Portland, Oregon, 24 June 2011.

metres above sea level.

We start in Section 2 with an overview of data-to-text generation and of related work in MT. In Section 3 we describe our comparable data and text resources and the pre-processing methods we apply to them. In Section 4 we provide an overview of our unsupervised learning task and of the methodology we have developed for it. We then describe our methods and results for sentence selection (Section 5) and sentence-level data selection (Section 6) in more detail. We finish with a discussion of our results and some conclusions (Section 7).

2 Background and Related Research

Work in data-to-text generation has involved a variety of 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 have included dynamic time-series data (such as meteorological or medical data) and static database entries (as in museum exhibits).

The following is an example of an input/output pair from the M-PIRO project (Androutsopoulos et al., 2005), where the input is a database record for a museum artifact, and the output is a description of the artifact:

```
creation-period=archaic-period,
current-location=Un-museum-Pennsylvania,
painting-techinique-used=red-figure-technique,
painted-by=Eucharides, creation-time=between
(500 year BC)(480 year BC)
```

Classical kylix

This exhibit is a kylix; it was created during the archaic period and was painted with the red figure technique by Eucharides. It dates from between 500 and 480 B.C. and currently it is in the University Museum of Pennsylvania.

While data and texts in the three example domains cited above do occur naturally, two factors mean they cannot be used directly as target corpora or training data for building data-to-text generation systems: one, most are not freely available to researchers (e.g. by simply being available on the Web), and two, more problematically, the correspondence between inputs and outputs is not as direct as it is, say, between a source language text and its translation. In general, naturally occurring resources of data and related texts are not 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 paper we look at the possibility of automatically identifying parallel data-text fragments from comparable corpora in the case of data-to-text generation from static database records. Such a parallel data-text resource could then be used to train an existing data-to-text generation system, or even to build a new statistical generator from scratch, e.g. using techniques from statistical MT (Belz and Kow, 2009).

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 and 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 subsentential fragments (Munteanu and Marcu, 2006) that are parallel within comparable corpora.

The latter approach is particularly relevant to our work. Munteanu and Marcu 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 that have only a small

number of 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 data-to-text generation situation is that in the former, sentence-aligned parallel resources exist and can be used as a starting point. E.g. Munteanu and Marcu use an existing parallel Romanian-English corpus to (automatically) create a lexicon which is then used in various ways in their method. In data-to-text generation we have no analogous resources to help us get started. The approach to data-text alignment described in this paper therefore uses no prior knowledge, and all our learning methods are unsupervised.

3 Data and Texts about British Hills

As a source of data, we use the Database of British Hills (BHDB) created by Chris Crocker,[1] version 11.3, which contains measurements and other information about 5,614 British hills. We add some information to the BHDB records by performing reverse geocoding via the Google Map API[2] 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 record.

On the text side, we use Wikipedia articles in the WikiProject British and Irish Hills (retrieved on 2009-11-09). At the time of retrieval there were 899 pages covered by this WikiProject, 242 of which were of quality category B or above.[3]

3.1 Aligning database entries with documents

Given that different hills can share the same name, and that the same hill can have several different names and spellings, matching up the data records in the BHDB with articles in Wikipedia is not entirely trivial. The method we use is to take a given hill's name from the BHDB record and to perform a search of Wikipedia with the hill's name as a search term, using the Mediawiki API. We then pair up the BHDB record with the Wikipedia article returned as the top search result.

We manually evaluated the data-text pairs matched by this method, scoring each pair good/unsure/bad. We found that 759 pairs out of 899 (the number of Wikipedia articles in the WikiProject British and Irish Hills at the time of retrieval), or 84.4%, were categorised 'good' (i.e. they had been matched correctly), a further 89 pairs (9.8%) were categorised 'unsure', and the remainder was a wrong match. This gave us a corpus of 759 correctly matched data record/text pairs to work with.

We randomly selected 20 of the data record/text pairs for use as a development set to optimise modules on, and another 20 pairs for use as a test set, for which we did not compute scores until the methods were finalised. We manually annotated the 40 texts in the development and test sets to mark up which subsets of the data and which text substrings correspond to each other for each sentence (indicating parallel fragments as shown at the bottom of Figure 2).

```
k-name v-name-Beacon_Fell
k-area v-area-Lakes:_S_Fells
k-height-metres v-height-metres-255
k-height-feet v-height-feet-837
k-feature v-feature-cairn
k-classification v-classification-WO
k-classification v-classification-Hu
k-locality v-locality-Skelwith
k-admin-area-level1 v-admin-area-level1-England
k-admin-area-level2 v-admin-area-level2-Cumbria
k-country v-country-United_Kingdom
```

Figure 1: Result of preprocessing BHDB record for Beacon Fell.

3.2 Pre-processing of data records and texts

Database records: We perform three kinds of preprocessing on the data fields of the BHDB database records: (1) deletion; (2) structure flattening, and (3) data conversion including the reverse geocoding mentioned above (the result of these preprocessing steps for the English hill Beacon Fell can be seen in Figure 1).

Furthermore, for each data field `key = value` we separate out key and value, prefixing the key with `k-` and the value with `v-key` (e.g. `v-area` and `k-area-Berkshire`). Each data field is thus con-

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

[2] http://code.google.com/apis/maps/

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

verted into two 'data tokens'.

Texts: For the texts, we first strip out Wikipedia mark-up to yield text-only versions. We then perform sentence splitting and tokenisation (with our own simple tools). Each text thus becomes a sequence of strings of 'text tokens'.

4 Task and Methodology Overview

Our aim is to automatically create aligned data-text resources where database records are paired with documents, and in each document, strings of word tokens are aligned with subsets of data tokens from the corresponding database record. The first two items shown in Figure 2 are the text of the Wikipedia article and the BHDB record about Black Chew Head (the latter cut down to the fields we actually use and supplemented by the administrative area information from reverse geocoding). The remainder of the figure shows fragments of text paired with subsets of data fields that could be extracted from the two comparable inputs.

How to get from a collection of texts and a separate but related collection of database records, to the parallel fragments shown at the bottom of Figure 2 is in essence the task we address. In order to do this automatically, we identify the following steps (the list includes, for the sake of completeness, the data record/document pairing and pre-processing methods from the previous section):

1. Identify comparable data and text resources and pair up individual data records and documents (Section 3).
2. Preprocess data and text, including e.g. tokenisation and sentence splitting (Section 3.2).
3. Select sentences that are likely to contain word strings that correspond to ('realise') any data fields (Section 5).
4. For each sentence selected in the previous step, select the subset of data tokens that are likely to be realised by the word strings in the sentence (Section 6).
5. Extract parallel fragments (future work).

5 Sentence Selection

The Wikipedia articles about British Hills in our corpus tend to have a lot of text in them for which the corresponding entry in BHDB contains no matching data. This is particularly true of longer articles about more well-known hills such as Ben Nevis. The article about the latter, for example, contains sections about the name's etymology, the geography, geology, climate and history, and even a section about the Ben Nevis Distillery and another about ships named after the hill, none of which the BHDB entry for Ben Nevis contains any data about. The task of sentence selection is to rule out such sections, and pick out those sentences that are likely to contain text that can be aligned with data. Using the example in Figure 2, the aim would be to select the first two sentences only.

Our sentence selection method consists of (i) estimating the strength of association between data and text tokens (Section 5.1); and (ii) selecting those sentences for further consideration that have sufficiently strong and/or numerous associations with data tokens (Section 5.2).

5.1 Computing positive and negative associations between data and text

We measure the strength of association between data tokens and text tokens using log-likelihood ratios which have been widely used for this sort of purpose (especially lexical association) since they were introduced to NLP (Dunning, 1993). They were e.g. used by Munteanu & Marcu (2006) to obtain a translation lexicon from word-aligned parallel texts.

We start by obtaining counts for the number of times each text token w co-occurs with each data token d, the number of times w occurs without d being present, the number of times d occurs without w, and finally, the number of times neither occurs. Co-occurrence here is at the document/data record level, i.e. a data token and a text token co-occur if they are present in the same document/data record pair (pairs as produced by the method described in Section 3). This allows us to compute log likelihood ratios for all data-token/text-token pairs, using one of the G^2 formulations from Moore (2004) which is shown in slightly different representation in Figure 3. The resulting G^2 scores tell us whether the frequency with which a data token d and a text token w co-occur deviates from that expected by chance.

If the G^2 score for a given (d, w) pair is greater than their joint probability $p(d)p(w)$, then the asso-

Wikipedia text:

Black Chew Head is the highest point (or county top) of Greater Manchester , and forms part of the Peak District , in northern England . Lying within the Saddleworth parish of the Metropolitan Borough of Oldham , close to Crowden , Derbyshire , it stands at a height of 542 metres above sea level . Black Chew Head is an outlying part of the Black Hill and overlooks the Chew Valley , which leads to the Dovestones Reservoir .

Entry from Database of British Hills:

name	area	height m	height ft	feature	classification	top	locality	adm_area1	adm_area2	country
Black Chew Head	Peak District	542	1778	fence	Dewey	Greater Manchester	Glossop	England	Derbyshire	UK

Parallel fragments:

name	area	top	adm_area1	adm_area2
Black Chew Head	Peak District	Greater Manchester	England	Derbyshire

height (m)
542

Black Chew Head is the highest point (or county top) of Greater Manchester , and forms part of the Peak District , in northern England .

it stands at a height of 542 metres above sea level .

Figure 2: Black Chew Head: Wikipedia article, entry in British Hills database (the part of it we use), and parallel fragments that could be extracted.

ciation is taken to be positive, i.e. w is likely to be part of 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.

Note that we use the notation G^2_+ below to denote a G^2 score which reflects a positive association.

5.2 Selecting sentences on the basis of association strength

In this step, we consider each sentence s in turn. We ignore those text tokens that have only negative associations with data tokens. For each of the remaining text tokens w^s in s we obtain $maxg2score(w^s)$, its highest G^2_+ score with any data token d in the set D of data tokens in the database record:

$$maxg2score(w^s) = \arg\max_{d \in D} G^2_+(d, w^s)$$

We then use these scores in two different ways to select sentences for further processing:

1. Thresholding: Select all sentences that have at least one text token w with $maxg2score(w) > t$, where t is a given threshold.

2. Greater-than-the-mean selection: Select all sentences whose mean $maxg2score$ (computed over all text tokens with positive association in the sentence) is greater than the mean of mean $maxg2score$s (computed over all sentences in the corpus).

The reason why we are not interested in negative associations in sentence selection is that we want to identify those sentences that are likely to contain a text fragment of interest (characterised by high positive association scores), and such sentences may well also contain material unlikely to be of interest (characterised by negative association scores).

5.3 Results

Table 1 shows the results for sentence selection, in terms of Precision, Recall and F_1 Scores. In addition to the two methods described in the preceding section, we computed two baselines. Baseline 1 selects just the first sentence, which yields a Precision of 1 and a Recall of 0.141 for the test set (0.241 for the development set), indicating that in the manually aligned data, the first sentence is always selected and that less than a quarter of sentences selected are first sentences. Baseline 2 selects all sentences which yields a Recall of 1 and a Precision of 0.318 for the test set (0.377 for the development set), indicating that around one third of all sentences were selected in the manually aligned data.

Greater-than-the-mean selection roughly evens out Recall and Precision scores, with an F_1 Score above both baselines. As for thresholded selection, applying thresholds $t < 10$ results in all sentences being selected (hence the same R/P/F_1 scores as for Baseline 2).[4] Very high thresholds (500+) result in

[4]This ties in with Moore's result confirming previous anec-

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

Figure 3: Formula for computing G^2 from Moore (2004) (N is the sample size).

	Development Set			Test Set		
Selection Method	P	R	F_1	P	R	F_1
1st sentence only (Baseline 1)	1.000	0.241	0.388	1.000	0.141	0.247
All sentences (Baseline 2)	0.377	1.000	0.548	0.318	1.000	0.483
Greater-than-the-mean selection	0.516	0.590	0.551	0.474	0.634	0.542
Thresholded selection $t = 60$	0.487	0.928	0.639	0.423	0.965	0.588

Table 1: Sentence selection results in terms of Precision, Recall and F_1 Score.

very high Precision ($> .90$) with Recall dropping below 0.15. In the table, we show just the threshold that achieved the highest F_1 Score on the development set ($t = 60$).

Selecting a threshold on the basis of highest F_1 Score (rather than, say, $F_{0.5}$) in our case means we are favouring Recall over Precision, the intuition being that at this stage it is more important not to lose sentences that are likely to have useful realisations in them (than it is to get rid of sentences that are not).

6 Data Selection

For data selection, the aim is to select, for each sentence remaining after sentence selection, the subset of data tokens that are realised by (some part of) the sentence. In terms of Figure 2, the aim would be to select for each of sentence 1 and 2 the data tokens which are shown next to the fragment(s) extracted from it at the bottom of Figure 2. Looked at another way, we want to get rid of any data tokens that are not likely to be realised by any part of the sentence they are paired with.

We preform sentence selection separately for each sentence s, obtaining the subset D_s of data tokens likely to be realised by s, in one of the following two ways:

1. Individual selection: Retain all and only those data tokens that have a sufficiently strong positive association with at least one text token w^s:

$$D_s = \left\{d \middle| \exists w^s (G^2_+(d, w^s) > t)\right\}$$

2. Pairwise selection: Consider each pair of key and value data tokens d^k_i, d^v_i that were originally derived from the same data field f_i. Retain all and only those pairs d^k_i, d^v_i where either d^k_i or d^v_i has a sufficiently strong association with at least one text token:

$$D_s = \left\{d^k_i, d^v_i \middle| \exists w^s_j \big(G^2_+(d^k_i, w^s_j) > t\big) \vee \exists w^s_m \big(G^2_+(d^v_i, w^s_m) > t\big)\right\}$$

Note that while previously each sentence in a text was associated with the same set of data tokens (the original complete set), after data selection each sentence is associated with its own set of data tokens which may be smaller than the original set.

If data selection produces an empty data token set D_s for a given sentence s, then s, along with its data token set D_s, are removed from the set of pairs of data token set and sentence.

We evaluate data selection for the baseline of selecting all sentences, and the above two methods in combination with different thresholds t. As the evaluation measure we use the Dice coefficient (a measure of set similarity), computed at the document level between (i) the union D of all sentence-level sets of data tokens selected by a given method and (ii) the corresponding reference data token set D^R, i.e. the set of data tokens in the manual annotations of the same text in the development/test data. Dice is defined as follows:

$$Dice(D, D^R) = \frac{2|D \cap D^R|}{|D| + |D^R|}$$

Table 6 shows results for the baseline and individual and pairwise data selection, on the development set

dotal evidence that G^2 scores above 10 are a reliable indication of significant association (Moore, 2004, p. 239).

		Sentence selection method			
		Greater-than-the-mean	Thresholded, $t = 60$	All-sentences	1st-sentence
Dev Set	All data tokens	0.666	0.666	0.666	0.666
	Individual selection	$t = 0$: 0.666	$t = 0$: 0.666	$t = 0$: 0.666	$t = 0$: 0.666
	Pairwise selection	$t = 19$: 0.706	$t = 18$: 0.709	$t = 18$: 0.717	$t = 1$: 0.697
Test Set	All data tokens	0.716	0.748	0.748	0.748
	Individual selection	$t = 0$: 0.716	$t = 0$: 0.748	$t = 0$: 0.748	$t = 0$: 0.748
	Pairwise selection	$t = 19$: 0.751	$t = 18$: 0.777	$t = 18$: 0.775	$t = 1$: 0.767

Table 2: Data selection results in terms of Dice coefficient. Results shown for data selection methods preceded by different sentence selection methods.

(top half of the table), and on the test set (bottom half). In each case we show results for the given data selection method applied after each of the four different sentence selection methods described in Section 5: greater-than-the-mean, thresholded with $t = 60$, and the first-sentence-only and all-sentences baselines (these index the columns).

Again, we optimised the two non-baseline methods on the development set, finding the best threshold t separately for each combination of a given data selection method with a given sentence selection method. This yielded the t values shown in the cells in the table.

Looking at the results, selecting data tokens individually (second row in each half of Table 6) cannot improve Dice scores compared to leaving the original data token set in place (first row); this is the case across all four sentence selection methods. The pairwise data selection method (third row) achieves the best results, although it does not appear to make a real difference whether or not sentence selection is applied prior to data selection.

7 Conclusion

In this paper we have reported our work to date on data-text alignment, a previously unexplored problem as far as we are aware. We looked at alignment of two comparable resources (one a collection of data records about British Hills, the other a collection of texts about British Hills) at the data record/document level, where our simple search-based method achieved an accuracy rate of 84%. Next we looked at alignment at the data record/sentence level. Here we obtained a best F_1 score of 0.588 for sentence selection and a best mean Dice score of 0.777 for data selection.

The best performing methods described here provide a good basis for further development of our parallel fragment extraction methods, in particular considering that the methods start from nothing and obtain all knowledge about data-text relations in a completely unsupervised way. Our results show that log likelihood ratios, which have been widely used for measuring lexical association, but were so far unproven for the data-text situation, can provide a strong basis for identifying associations between data and text.

References

I. Androutsopoulos, S. Kallonis, and V. Karkaletsis. 2005. Exploiting owl ontologies in the multilingual generation of object descriptions. In *Proceedings of the 10th European Workshop on Natural Language Generationf (ENLG'05)*, pages 150–155.

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.

E. Briscoe, J. Carroll, and J. Graham. 2006. The second release of the rasp system. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*.

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

A. Isard, J. Oberlander, I. Androutsopoulos, and C. Matheson. 2003. Speaking the users' languages. *IEEE Intelligent Systems Magazine: Special Issue "Advances in Natural Language Processing*, 18(1):40–45.

Robert C. Moore. 2004. 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.