Monolingual Data Optimisation for Bootstrapping SMT Engines

Jie Jiang,[†] **Andy Way,**[†] **Nelson Ng,**[‡] **Rejwanul Haque,**[†] **Mike Dillinger,**[‡] **and Jun Lu**[‡]
[†]Applied Language Solutions, Delph, OL3 5FZ, UK
[‡]eBay Inc.
{jie.jiang,andy.way}@appliedlanguage.com, nng@ebay.com,
rejwanul.haque@appliedlanguage.com, {micdillinger,jlu}@ebay.com

Abstract

Content localisation via machine translation (MT) is a sine qua non, especially for international online business. While most applications utilise rule-based solutions due to the lack of suitable in-domain parallel corpora for statistical MT (SMT) training, in this paper we investigate the possibility of applying SMT where huge amounts of monolingual content only are available. We describe a case study where an analysis of a very large amount of monolingual online trading data from eBay is conducted by ALS with a view to reducing this corpus to the most representative sample in order to ensure the widest possible coverage of the total data set. Furthermore, minimal yet optimal sets of sentences/words/terms are selected for generation of initial translation units for future SMT system-building.

1 Introduction

On many occasions clients approach machine translation (MT) providers knowing that they do not have parallel data that can be used 'as is' to train statistical MT (SMT) engines. The typical approach in such circumstances is to mine the Web for data that is as 'close as possible' to the specific use-case of the client that can be used – after cleaning – as parallel training data (cf. Pecina et al. (2011)).

However, not all clients are the same, and solutions such as the above may not be appropriate in all cases. Sometimes clients have hundreds of millions of sentences of monolingual data only, which is already publicly available. Searching on the Web for 'similar' data would be meaningless, as one would typically pick up that self-same data in a web crawl.

In this paper, we describe the engagement of the language technology (LT) group at Applied Language Solutions (ALS) by eBay, the world's largest multinational online trading company, to research the feasibility of providing multilingual content from their monolingual data using MT technologies in order to facilitate a multilingual cross-border trading solution for eBay. ALS were provided with a large sample of eBay's English data – mostly user-generated content – with a view to recommending which parts of that data set were most representative of the data as a whole, and which could then be set aside for human translation so as to seed an initial parallel data set for SMT engine-building.

With the vast amount of data provided by eBay, it was essential to analyse the content prior to any further text processing, given the strong probability of a large variety of domains and text genres in the data set as a whole. We describe the corpus in detail in Section 2, but essentially the data comprised 34 separate eBay categories, with each separate data item in each category consisting of 7 fields. We needed to discover the 'closeness' of these categories, to try to reduce the number of MT engines that we would recommend be built for eBay. Once we had established this, we were required to select the minimum number of monolingual sentences to create Translation Units (TUs), which once translated (either (i) completely by hand, or (ii) via MT seeded with some manually translated data, followed by post-editing) could be used to train the engines to translate the remainder of the data, together with any new incoming

Tag Name	Total No. Sentences	No. Unique Sentences	Sentence Duplicates	No. Words	Vocabulary Size
Item Title	1,016,364	1,001,169	1.5%	8,960,683	87,580
Item Subtitle	113,007	24,040	78.73%	772,599	8,042
Item Description	32,193,213	6,194,681	80.76%	437,020,245	277,252
Payment Instructions	1,296,336	142,725	88.99%	17,663,422	14,520
Refund As	741,956	6	100%	1,399,393	13
Return Within	741,956	9	100%	1,483,912	9
Refund Details	1,447,691	108,778	92.49%	24,131,176	12,019
Totals:	37,550,523	7,471,408		491,431,430	399,435

Table 1: Monolingual Data Corpus Statistics

text. In brief, we performed monolingual data analysis on vocabulary overlap, and monolingual data clustering to find the optimal yet smallest number of domains. Finally, we extracted terminology for each cluster and then used a term-based selection process to select the optimal TUs for translation so as to provide SMT training data.

In the following sections, we introduce each of these tasks and procedures that we taken to solve the related questions. While all the experiments were carried out on English data, the methods used are generic to any language, so it is easily extensible to other eBay source-language material.

The rest of the paper is organised as follows. Section 2 describes the half a billion running words of mostly user-generated eBay content (in English) that the LT group in ALS had to deal with, with analysis performed on several levels. Section 3 describes the monolingual data clustering methods used by ALS to come up with the ideal source-side of a parallel corpus, source-language lexical and terminological data that post-translation would be optimal for SMT training that we outline in Section 4. We conclude in Section 5, and provide ways in which this analysis may be extended, in terms of other eBay user-generated data, and/or for languages other than English, or by using other techniques.

2 Monolingual Data Analysis

The monolingual data provided by eBay has 34 categories, namely: Antiques, Art, Baby, Books, Business & Industrial, Cameras & Photo, Cell Phones & PDAs, Clothing Shoes & Accessories, Coins & Paper Money, Collectibles, Computers & Networking, Crafts, Dolls & Bears, DVDs & Movies, Electronics, Entertainment Memorabilia, Everything Else, Gift Cards & Coupons, Health & Beauty, Home & Garden, Jewellery & Watches, Music, Musical Instruments, Pet Supplies, Pottery & Glass, Real Estate, Speciality Services, Sporting Goods, Sports Memorabilia Cards & Fan Shop, Stamps, Tickets, Toys & Hobbies, Travel and Video Games.

Each category contains different amounts of items. Each item contains 7 different fields (labelled with different tags), namely Item Title, Item Subtitle, Item Description, Payment Instructions, Refund As, Refund Within and Refund Details. Note that the vast majority of the data comprises user-generated content, although the three ‘Refund’ fields contain a lot of boilerplate material, as we will demonstrate later.

In the following sections, we firstly describe the data pre-processing performed, followed by our detailed analysis of the monolingual data, in order to compute the closeness of the different eBay categories, and the different fields within each item.

2.1 Data pre-processing

The total size of the original English data was 34.8GB in XML format. The following steps were performed to pre-process the data:

- Extraction of pure text content by stripping tags and Javascript. The text of all items was labelled with the corresponding category and tag information for further processing downstream.
- Filtering of non-English material using English dictionaries, since some of the data con-

Tag Name	Vocabulary Size before Noise Reduction	Vocabulary Size after Noise Reduction
Item Title	87,580	38,352
Item Subtitle	8,042	6,885
Item Description	277,252	71,820
Payment Instructions	14,520	9,518
Refund As	13	13
Return Within	9	9
Refund Details	12,019	8,292
Totals:	399,435	134,889

Table 2: Monolingual Data Corpus Statistics after Noise Reduction

tained characters in German, Chinese, Bulgarian, Ukrainian, Russian, French, Arabic, Sanskrit, Spanish, Greek, Armenian, French, Polish and Japanese.

- Segmentation of sentences via a set of regular expressions, and then typical MT corpus pre-processing including tokenisation and lowercasing with the Moses toolkit (Koehn et al., 2007).

2.2 Corpus statistics

After pre-processing, the statistics on the eBay data set are provided in Table 1. It is easy to see that we are dealing with huge data sizes here. From the initial 34.8GB of data, after cleaning, we see there are 37.5M sentences, the vast majority (85.7%) included in Item Descriptions. With MT as the end-task in mind, it is clear that this will only be practical if smaller samples of this overall data set can be used as training data for the language-pairs of interest to eBay. Things become much more realistic when we see in the next column that the number of unique sentences is around 7.5M; that is, 80.2% of the data is found more than once in the overall data set. It would have been reasonable to assume that the Titles, Subtitles and Descriptions are particular to the Items themselves, while Payment Instructions, Refund and Return Details are similar for each item. However, as we can see, even Item Subtitles (78.7%) and Descriptions (80.8%) contain a huge number of duplicate sentences. Only the Item Titles themselves seem to be unique.

Even so, training MT engines with 7.5M sentence-pairs is non-trivial, but of course here the data we have is only monolingual. Accordingly, for MT deployment in eBay to be practical, samples from this smaller data set still need to be selected. As we move to words, the figures become even more staggering, with a total of almost half a billion words in the set. However, only around 400K (or 0.0008%) of these words are unique. Many of these are real names, places, etc., many of which will not need to be translated, with quite a few typos contained therein to boot, which can be cleaned up prior to further processing. Accordingly, we automatically removed all typos and entries not contained in two English dictionaries and recalculated the statistics in Table 1, leading to the new numbers for ‘Vocabulary Size’ in Table 2.

As can be seen, the removal of the different types of noise significantly reduces the amount of vocabulary items that need to be handled. Overall, numbers decrease by 264,546, or 66%. As expected, the biggest savings are to be had for Item Descriptions (205,432 words, or 74%), but signifant reductions are seen for all other major categories: Item Titles (49228 words, or 56%), Item Subtitles (1157 words, or 14%), Payment Instructions (5002 words, or 34%), and Refund Details (3727 words, or 31%).

All of the above gives us cause for optimism, as it is clear that for multinational multilingual companies such as eBay, automation is the only way in which data sizes of the amounts shown above can be handled.

2.3 Vocabulary overlap

To investigate the closeness between each of the 34 eBay categories, we calculated the vocabulary over-

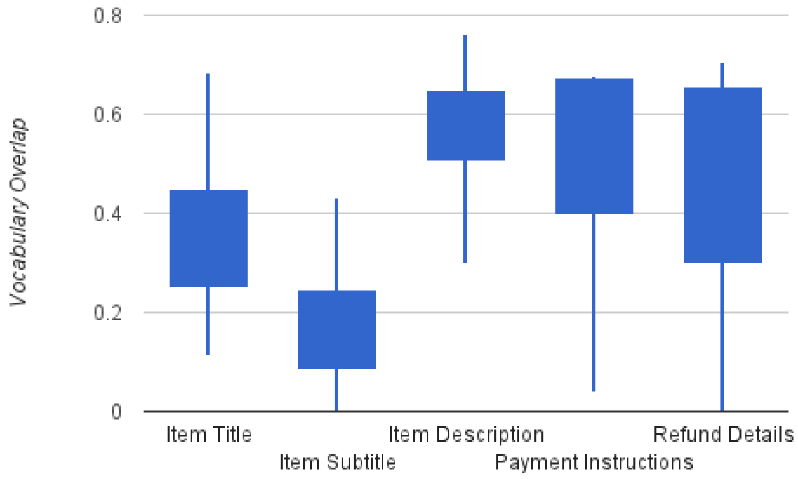

Figure 1: Vocabulary Overlap per tag between different eBay categories

lap across all categories, where the denominator is the super-set of both dictionaries. Clearly such a grid contains far too many entries to be inserted in this paper, so instead we illustrate only the maximum, minimum and standard derivations between all one-to-one vocabulary overlaps between each of the categories. The results are shown in Figure 1, and separated by different tag names. Note that we ignored Refund As and Returned Within since they are trivial to analyse with such a tiny vocabulary size.

The general observations are as follows:

- *Item Titles*: in general, vocabulary overlap is on the low side here, varying from around 11% (Real Estate ∩ Cell Phones & PDAs) to 64% (Entertainment Memorabilia ∩ Music). Some of this may be due to data sparseness, but more generally may be attributed to the free-form input in Item Titles. Some categories tend to show that they have low overlap with most other categories (e.g. Real Estate, Tickets), while others show relatively high overlaps across the board (e.g. Collectibles, Crafts, Toys & Hobbies). Generally speaking, average overlaps seem to be around 35% for this sub-part of the data.

- *Item Subtitles*: as for Titles, vocabulary overlap for Item Subtitles is low, on average around 1520%. The range of overlap varies from less than 1% (Entertainment Memorabilia ∩ Real Estate) to 43% (Electronics ∩ Cameras & Photo). Some of these very low overlaps are due to data sparseness, as this field appears to be optional, so quite low numbers of instances can be seen in the data. Overlaps are low across the board for some categories (Real Estate, Dolls & Bears, Gift Cards & Coupons, Specialty Services, Sports Memorabilia, Stamps), while for others they are relatively high, on average (Cameras & Photo, Computers & Networking, Music, Toys & Hobbies).

- *Item Descriptions*: here, for the sub-part of the Items that comprises by far the largest amount of the data, vocabulary overlap varies from around 35% (Tickets ∩ Books) to 73% (Cameras & Photo ∩ Computers & Networking). More generally overlaps of around 60% are seen. Some categories tend to show that they have low overlap with most other categories (e.g. Tickets, Real Estate), while others show relatively high overlaps across the board (such as Musical Instruments, Home & Garden, Collectibles and Cameras & Photo).

- *Payment Instructions*: The percentage of vocabulary overlap for this sub-part of the data is in general very high, with average overlaps of over 60%. These range from less than 4% (Real Estate ∩ Pottery) to 68% (Computers & Networking ∩ Cameras & Photo). Again, low average scores can be seen for certain categories (Real Estate, Tickets), while others have higher than average vocabulary overlaps (Toys & Hobbies, Computers & Networking).

- *Refund Details*: Again, the average vocabulary overlaps for this sub-part of the data are reasonably high, around 50% overall. Values range from 0% (Real Estate ∩ Any other category) to 71% (Cameras & Photo ∩ Electronics). As before, scores which are on the low side are seen for certain categories (Specialty Services, Gift Cards & Coupons), with others having high average overlaps (Cameras & Photo, Computers & Networking, Electronics).

In sum, vocabulary overlaps for Item Titles and Item Subtitles are on the low side, whereas high average overlaps are seen for Item Descriptions, Payment Instructions and Refund Details. Given the large overlaps at sentential level for the latter two sub-parts, these will need to be translated only once to ensure that the vast majority of future cases in these data fields will be covered and accurately translated. Where Item Titles and Item Subtitles are concerned, these will largely be covered by accurate translation of the termbanks extracted from monolingual data.

Since Item Descriptions comprise by far the largest portion of the data, the vocabulary overlap seen there is encouraging when we consider the training data samples extracted in the next section. Thus we will focus on Item Descriptions where clustering and data selection is concerned.

3 Monolingual Data Clustering

The aim in this section is to find an optimal number of MT engines to translate the monolingual data once eBay's multilingual cross-border trading solution goes live. The obvious solutions are either to use one single generic engine or 34 domain-specific engines, but these are unlikely to be the best ways forward for eBay. Accordingly, we employ data clustering techniques to identify optimal clusters based on the 34 categories for SMT engine-building.

3.1 Clustering features and algorithms

We used three different features to perform clustering on the eBay monolingual data, namely:

1. TF-IDF (Spärck Jones, 1972; Salton and McGill, 1983; Salton et al., 1983; Salton and Buckley, 1988; Wu et al., 2008),
2. Language Model (LM) Perplexity (Ponte and Croft, 1998; Song and Croft, 1999; Lv and Zhai, 2009; Manning et al., 2009; Büttcher et al., 2010),
3. Dice Coefficient (Dice, 1945; van Rijsbergen, 1979; Kondrak et al., 2003).

The problem we are confronted with here is an instance of unsupervised learning, where the data is essentially unstructured, with no annotations to guide the machine-learning process. In our case, the chosen algorithm is provided with the data alone, and has to learn some basic characteristics of that data via distributional patterns, and in this section, clustering.

We employed the following two approaches for clustering:

1. Hierarchical clustering (Press et al., 2007a; Hastie et al., 2009),
2. K-Means clustering (Kanungo et al., 2002; Press et al., 2007b).

We tried all combinations of features and algorithms, and our general findings were as follows:

- Different features tend to produce different clusters,
- Different algorithms tend to produce similar clusters. However, K-Means with a random start point tend to produce clusters with small differences.

Since our aim was to find optimal clusters for MT engine building, we chose the LM perplexity feature instead of the other two, because it is closely related to SMT performance: the lower the LM perplexity, the better the MT engine's performance with respect to translation quality, as has been widely reported (e.g. (Eck et al., 2004; Foster & Kuhn, 2007)). We also chose hierarchical clustering as we want the clustering results to be stable as opposed to changing over time.

Specifically, LM perplexity is calculated by dividing the data in each of the categories into independent training and testing sections, with random sampling from corresponding items. Language models are built on the training data, and then both within- and cross-category LM perplexities are calculated on the test data. All the perplexity scores are normalized on the in-domain perplexity score, and these are used for the distance measure of the clustering procedure in Section 3.3. We do this so that the scores are comparable across categories; this needs to be done given the varying amounts of data in each of the eBay categories. Lest there be any misunderstanding, this normalization has nothing to do with the calculation of the perplexity scores *per se*.

3.2 Optimal number of clusters

For both Hierarchical clustering and K-Means clustering, we need to determine the optimal number of clusters. For the LM Perplexity feature, average in-cluster LM Perplexity (PPL_{avg}) is used, as in formula (1):

$$PPL_{avg} = \frac{\sum_i \frac{\sum_{j,k} d(c_j,c_k)}{M_i}}{N} \qquad (1)$$

where N is the number of clusters, i is the index for cluster i, M_i is the number of categories in cluster i; j and k are the indexes for categories in cluster i,and c_j and c_k are categories j and k in cluster i, $d(c_j, c_k)$ are the perplexity scores calculated by c_j text with LM model built via random sampling on c_k.

To balance the number of MT systems required and LM perplexities, the dynamics of this objective function indicate the benefits of overall LM perplexities by increasing/decreasing the total number of clusters. Accordingly, we select the number of clusters which has the biggest drop in value of the objective function.

3.3 Clustering results

The hierarchical clustering results using the LM perplexity feature on Item Description data is shown in Figure 2. Note that different clusters can be obtained with varying distance thresholds. Therefore, PPL_{avg} in formula (1) is used to determine the optimal number of clusters. PPL_{avg} and its dynamics with respect to the change in number of clusters are illustrated in Figure 3. The blue curve shows the average in-cluster LM perplexities (PPL_{avg}), and the red curve depicts the different changes in cluster numbers ($PPL_{avg}(N-1) - PPL_{avg}(N)$ at point N).

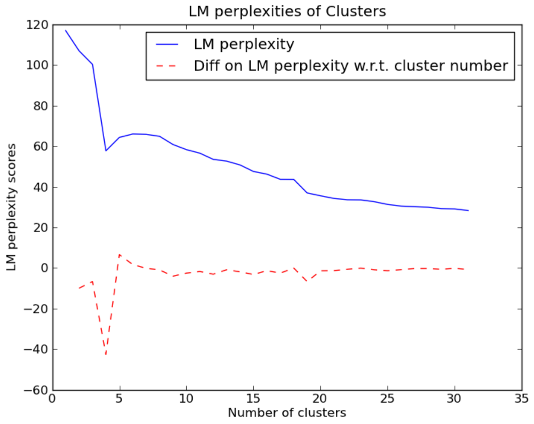

Figure 3: PPL_{avg} of different number of clusters in Hierarchical Clustering

Figure 3 clearly shows that 4 clusters appears to be the best trade-off between average LM perplexity and number of clusters, as there is a big drop in PPL_{avg} from three clusters to four, and results do not change dramatically thereafter.

Accordingly, the optimal 4 clusters which result from hierarchical clustering for Item Descriptions are as follows:

- Cluster 1: “Baby”, “Business Industrial”, “Cameras Photo”, “Cell Phones PDAs”, “Computers Networking”, “Dolls Bears”, “Electronics”, “Health Beauty”, “Home Garden”, “Musical Instruments”, “Pet Supplies”, “Sporting Goods”, “Sports Mem Cards Fan Shop”, “Tickets”, “Toys Hobbies”, “Travel”, “Video Games”

- Cluster 2: “Antiques”, “Art”, “Books”, “Clothing Shoes Accessories”, “Coins Paper Money”, “Collectibles”, “Crafts”, “DVDs Movies”, “Entertainment Memorabilia”, “Everything Else”, “Jewelry Watches”, “Music”, “Pottery Glass”, “Specialty Services”, “Stamps”

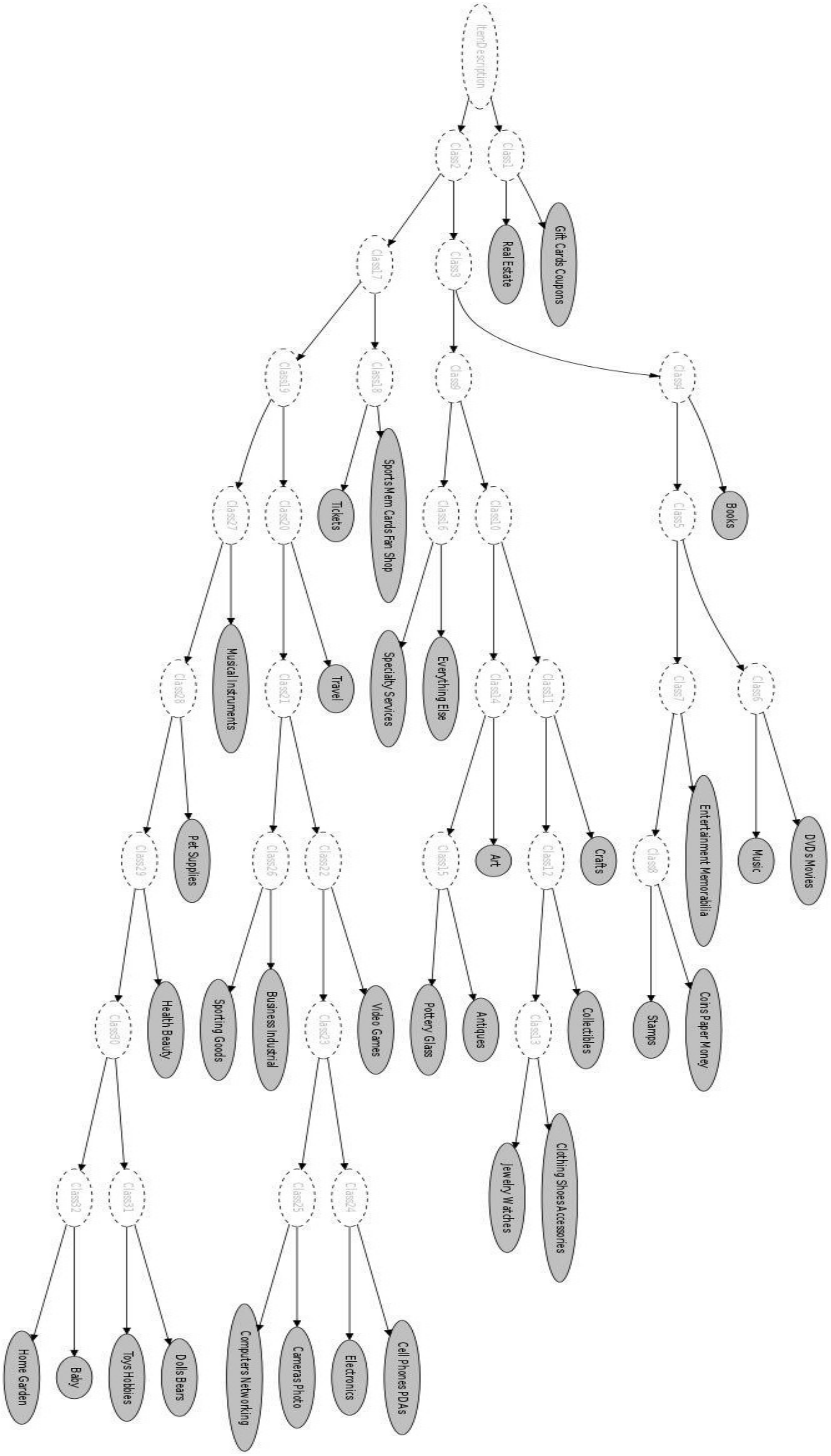

Figure 2: Tree structure of categories based on Hierarchical Clustering

- Cluster 3: “Gift Cards Coupons”
- Cluster 4: “Real Estate”

Note that two quite specific categories ‘Gift Cards & Coupons’ and ‘Real Estate’ end up in separate clusters of their own. Clearly the text content in these two categories is quite different from other categories – as we saw earlier in Section 2.3 – and should be treated differently with specific MT engines, while all other categories can be handled with two MT engines with much wider coverage in terms of domains.

With this 4-cluster information, we are able to select optimal monolingual sentence sets to seed MT corpus building, which we describe next.

4 Optimal Data for SMT Training

The results in the previous two sections are applied to find the optimal data collections for consideration as the source-side of SMT training data. As in the previous sections, we focus primarily on the selection of training data for SMT engines for Item Description data.

The first step is to extract terminology sets for each of the categories for Item Description data. TF-IDF is used to select those terms which have higher values than a given threshold. We select the threshold heuristically based on the score distribution, specifically choosing the point at which there is a significant drop in TF-IDF scores. To be more precise, we cut off when the number of terms gained at the current point is no more than X% of the previously accumulated terms. X is different across categories, and typically it varies from around 10–20%.

Then a term-based TU selection process is carried out independently for each cluster. During the selection process, we plotted the relationship between the selected words and term/vocabulary coverage. We then use these statistics to determine the optimal selection point.

We plot the relationship in a graphical representation as follows:

1. Term coverage: percentage of terms covered,
2. Vocabulary coverage: percentage of vocabulary entries covered,

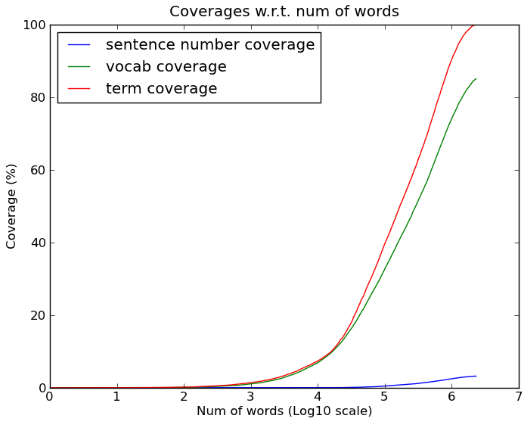

Figure 4: Data Selection Statistics for Cluster 1

3. Sentence number coverage: percentage of sentences covered.

Based on the statistics, we selected the optimal TUs where an increase in the number of words does not significantly benefit the term coverage. We use a heuristic threshold of 0.5 (on the log-scale) as the minimum increase to select the cut-off point to determine the number of words to be selected.

The selection graph of Cluster 1 is shown in Figure 4; the other three clusters have similar graphs, and are thus omitted for reasons of space. Actually the selection process is guided by the red curve in the Figure, which indicates the percentage of terms covered by the selected TUs. The optimal threshold is where term coverage is 91%.

By applying the heuristic threshold for all four clusters, we finally selected 4 different sets of TUs for the source-side of SMT training data, collected together in Table 3.

As Table 3 illustrates, across all 4 clusters for Item Descriptions, the optimal configuration only requires 2.17M words, which is less than 0.5% of the total initial set of 437 million words in Table 1.[1] Of these 2.17M words, 91% of the terms and 69% of the vocabulary are already covered by that data, showing that to obtain full terminology and vocabulary coverage, just under 4K more terms and 49K extra vocabulary items need to be translated in addition to

[1] This estimate is on the optimistic side, as it assumes that the target language has a comparable vocabulary size to English, which is not true for many (or most) languages.

No. Words	**Term Coverage**	**No. Terms Uncovered**	**Vocab Coverage**	**No. Vocab Uncovered**	**Total Words**
2,176,009	91%	3,959	68.7%	49,022	2,228,990

Table 3: Summary of Optimal TU selection

the source-side sentential data. This reduction in TU selection massively reduces the amount of data to be considered during the MT system-building process.

5 Conclusions and Further Work

In this paper, the LT group from ALS, a well-known translation service provider, analysed a huge sample of monolingual content sampled from eBay, the largest multinational online trading company. For 34 different categories with 7 different fields per Item, we calculated the sentence duplication and vocabulary overlaps to show that for most of the data, MT can be a suitable solution for content localisation where training data does not exist *a priori*. We then obtained the optimal number of clusters via monolingual data clustering, and then selected optimal sentences/words/terms that can be used to create an initial parallel corpus to train SMT engines for each cluster. Compared with the original huge data size, we demonstrated that only a very small amount of the total words need to be translated in the optimal training corpus to ensure maximum coverage.

What is absolutely certain is that for multinational companies like eBay who are interested in a multilingual solution, automation is key; handling data sizes of the magnitude that ALS had to deal with in this paper would be unthinkable by humans alone.

As far as extensions to this work are concerned, similar analysis could be performed for source languages other than English, since all of the techniques applied to the eBay English data can be easily extended to other languages with only small modifications. Meanwhile, further eBay user-generated content, such as buyer-seller interactions, and review guides and catalogue information are interesting data for further analysis.

In addition, we would like to see how cross-lingual information retrieval techniques (e.g. (Snover et al., 2008)) to automatically select parallel data compare to our method. Furthermore, given the similarity of the use-case described here to patent translation, approaches to domain-adaptation (e.g. (Banerjee et al., 2011)) or multi-task learning for patent translation (e.g. (Tinsley et al., 2010; Ceausu et al., 2011; Wäschle & Riezler, 2012) might be applicable to the e-commerce domain.

We are also interested in deeper levels of analysis arising from the study carried out in this paper. For example, we could have used the three clustering metrics together to seek further corroboration of the clusters that we ended up with, or used other techniques altogether (cf. (Mandal et al., 2008; Biçici & Yuret, 2011)). Finally, of course, the next stage is to build, apply, and evaluate the SMT engines constructed on the basis of the recommendations provided here on new, unseen eBay data, and compare the results of other method against these other possible approaches.

Acknowledgments

Many thanks to the anonymous reviewers, especially for the suggestions regarding related work.

References

Banerjee, P., S. Naskar, J. Roturier, A. Way and Josef van Genabith. 2011. Domain Adaptation in Statistical Machine Translation of User-Forum Data using Component Level Mixture Modelling. In *Proceedings of Machine Translation Summit XIII*, Xiamen, China, pp. 285–292.

Biçici, E. and D. Yuret. 2011. Instance selection for machine translation using feature decay algorithms. In *Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT '11*, Edinburgh, Scotland, pp. 272–283.

Büttcher, S., C. Clarke and G. Cormack. 2010. *Information Retrieval: Implementing and Evaluating Search Engines*. Cambridge MA: MIT Press, pp. 289–291.

Ceausu, A., J. Tinsley, J. Zhang and A. Way. 2011. Experiments on domain adaptation for patent machine translation in the PLuTO project. In *Proceedings of the 15th Annual Meeting of the European Association for Machine Translation*, Leuven, Belgium, pp. 21–28.

Dice, L. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology* 26(3): 297–302.

Eck, M., S. Vogel and A. Waibel. 2004. Language Model Adaptation for Statistical Machine Translation based on Information Retrieval. In *LREC-2004: Fourth International Conference on Language Resources and Evaluation, Proceedings*, Lisbon, Portugal, pp. 327–330.

Foster, G. and R. Kuhn. 2007. Mixture-model adaptation for SMT. In *ACL 2007: proceedings of the Second Workshop on Statistical Machine Translation*, Prague, Czech Republic, pp. 128–135.

Hastie, T., R. Tibshirani and J. Friedman. 2009. *The Elements of Statistical Learning* (2nd ed.). New York: Springer, pp. 520–527.

Kanungo, T., D. Mount, N. Netanyahu, C. Piatko, R. Silverman and A. Wu. 2002. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. Pattern Analysis and Machine Intelligence* **24**: 881–892.

Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin and E. Herbst. 2007. Moses: open source toolkit for statistical machine translation. In *ACL 2007: Proceedings of Demo and Poster Sessions*, Prague, Czech Republic, pp. 177–180.

Kondrak, G., D. Marcu and K. Knight. 2003. Cognates Can Improve Statistical Translation Models. In *Proceedings of HLT-NAACL 2003: Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Edmonton, Canada, pp. 46–48.

Lv, Y. and C-Z. Zhai. 2009. Positional Language Models for Information Retrieval. in *Proceedings of the 32nd International ACM SIGIR conference on Research and Development in Information Retrieval (SIGIR)*, Boston MA., pp. 299–306.

Mandal, A., D. Vergyri, W. Wang, J. Zheng, A. Stolcke, G. Tur, D. Hakkani-Tür and N. Ayan. 2008. Efficient Data Selection for Machine Translation. In *Proceedings of the 2008 IEEE Workshop on Spoken Language Technology*, Goa, India, pp. 261–264.

Manning, C., P. Raghavan and H. Schütze. 2009. An Introduction to Information Retrieval. In C. Manning & H. Schütze (eds.) *Foundations of Statistical Natural Language Processing*, Cambridge University Press, pp. 237–240.

Pecina, P., A. Toral, A. Way, P. Prokopidis and V. Papavassiliou. 2011. Towards Using Web-Crawled Data for Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 15th Annual Meeting of the European Association for Machine Translation*, Leuven, Belgium, pp. 297–304.

Ponte, J. and B. Croft. 1998. A Language Modeling Approach to Information Retrieval. In *SIGIR '98, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, pp. 275–281.

Press, W., S. Teukolsky, W. Vetterling and B. Flannery. 2007a. Section 16.4. Hierarchical Clustering by Phylogenetic Trees. *Numerical Recipes: The Art of Scientific Computing* (3rd ed.). Cambridge: Cambridge University Press.

van Rijsbergen, K. 1979. *Information Retrieval*. London: Butterworths.

Salton, G. and C. Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management* **24**(5): 513–523.

Salton, G., E. Fox and H. Wu. 1983. Extended Boolean Information Retrieval. *Communications of the ACM* **26**(11): 1022–1036.

Salton, G. and M. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Song, F. and B. Croft. 1999. A General Language Model for Information Retrieval. In *Proceedings of the 1999 ACM SIGIR Conference on Research and Development in Information Retrieval*. Berkeley CA., pp. 279–280.

Snover, M., B. Dorr and R. Schwartz. 2008. Language and translation model adaptation using comparable corpora. In *EMNLP 2008: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, Honolulu, HI, USA, pp. 857–866.

Spärck Jones, K. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation* **28**(1): 11–21.

Tinsley, J., A. Way and P. Sheridan. 2010. PLuTO: MT for Online Patent Translation. In *AMTA 2010: The Ninth Conference of the Association for Machine Translation in the Americas, Proceedings*, Denver, CO., pp. 435–442.

Wäschle, K. and S. Riezler. 2012. Structural and topical dimensions in multi-task patent translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, EACL-12*, Avignon, France, pp. 818–828.

Wu, H., R. Luk, K. Wong and K. Kwok 2008. Interpreting TF–IDF term weights as making relevance decisions. *ACM Transactions on Information Systems* **26**(3): 1–37.