Supplementary Material: Learning Semantic Correspondences in Technical Documentation

Kyle Richardson and Jonas Kuhn Institute of Natural Language Processing University of Stuttgart {kyle, jonas}@ims.uni-stuttgart.de

1 Dataset Information

We report additional details about our dataset collection.

1.1 Credits

Standard Library Documentation Figure 1 shows additional details of the different standard datasets, including pointers to the original source of the documentation, the standard library version numbers, and other details. To our knowledge, none of these datasets, excluding the Java Standard Library set from (Deng and Chrupała, 2014), have been used for the types of NLP experiments that we describe in the main paper.

Technical Manuals We also show information about the Unix dataset first reported in (Richardson and Kuhn, 2014). We are the first to report semantic parsing results using this dataset, since their original paper is only about the data resource.

1.2 Java Documentation Comparison

While we use the data (including the original splits) from Deng and Chrupała (2014) and apply some of the same baseline models, we made a few changes to their original evaluation, thus making our results not directly comparable. First, they train their models on the full documentation string, which we found to not have any advantage over just using the first sentence in the description (as we do for all other datasets). Secondly, they evaluate by ranking each example over a subset of the components in the validation/test splits, whereas we rank over all known components in the overall API (as also done in all other cases). We also made small differences changes to how tokenization is done in order to match the other datasets.

Despite these differences, as well as differences in the implementations of the baseline translation models, we were still able to reproduce (and in some cases improve) their original results.

1.3 Non-English Datasets

The documentation collections in languages other than English come exclusively from the PHP documentation set. Currently, these languages include: French, Spanish, Japanese, Russian, Turkish, and German (see paper for more details). More information about the translations, which are derived from the English original documentation set, can be found here: http://doc.php.net/tutorial/

1.4 Document Feature Information

Figure 1. shows information about the additional document-level features associated with each dataset.

Class information Some documentation sets include assertions about related functions or utilities, in the form of see also sections or links to other parts of the API. Such information can also be found in quick reference manuals or language cheat sheets available online, as well as from html structure. This information is used to define features in our discriminative model (see details in main paper).

Examples are shown in Figure 2. Since the datasets differ in terms of resources, the features used for each dataset are shown in gray.

Parameter descriptions In many datasets, the function documentation include additional textual descriptions of the function parameters. For the baseline translation models, these can add these fragmented pairs to the parallel training data. We also use this information as features in our discriminative model.

Return descriptions Similarly, some documentation also contains textual descriptions of return values, which can be used in the same way as described above.

				Docu	ment Fe	atures	
Dataset	Version	Source(s)		class info.	para	m. retu	n section
Java	SE 6.0	(Deng and Chrupała, 2014)	main data	see-also	~	✓	 Image: A second s
Ruby	2.3.0	ruby-doc.org/core-2.3.0/ ruby-doc.org/stdlib-2.3.0/	main data main data	fun. links	\checkmark	×	×
PHP	3.0	php.net/download-docs.php	main data	see-also	<	<	\checkmark
Python	2.7.11	<pre>docs.python.org/2.7/library/ docs.python.org/2/reference/</pre>	main data background	× html	× _	× _	✓ _
		docs.python.org/2.7/library/(numpy)	num. library	html	~	✓	X
Elian	25.1	gnu.org/software/emacs/manual wikemacs.org/wiki/Emacs_Lisp_Cheat_Sheet	main data background	× html	✓ _	× -	✓ _
Elisp	25.1	github.com/magnars/s.el(<i>s.el</i>) github.com/magnars/dash.el(<i>dash.el</i>)	string library list library	html html	××	××	××
Haskell	4.8.1	hackage.haskell.org/package/base-4.8.1.	0 main data	X	\times	×	 Image: A second s
		clojure.org/api/api clojuredocs.org/ clojure.org/api/cheatsheet	main data main data background	× see-also html	××	××	×
Clojure	1.7	<pre>github.com/weavejester/medley(medley) github.com/Raynes/fs(fs) github.com/ztellman/gloss(gloss) github.com/clj-time/clj-time(clj-time)</pre>	fun. library file-sys library byte library time library	× × ×	× × ×	× × ×	× × ×
С	2.24	gnu.org/software/libc/manual/ en.cppreference.com/w/c	main data main data	× html	× ✓	×	√ √
MIT Scheme	9.2	https://www.gnu.org/software/mit-scheme	main data	html	×	Х	×
Unix	_	(Richardson and Kuhn, 2014)	main data	see-also	✓	Х	√

Figure 1: Further details of our corpus collection, including any background resources (**background**) or third party libraries (shown under double line) that were used. The last column shows additional document features used in our experiments. *Class* refers to information about general classes of functions, and *param* and *return* specify if the additional textual description of parameter values and return values (respectively) are included.

Language	Example Classes
Ruby	<pre>{logger.info, logger.warn, logger.fatal, logger.debug, }</pre>
Elisp	<pre>{sin, cos, tan, asin, acos, atan, exp, log, log10, }</pre>
Unix	<pre>{iotop,iosnoop,iopattern,iopending, }</pre>
PHP	<pre>{ps_close_image, ps_open_image_file, ps_place_image, }</pre>
С	{UINT8_MAX UINT16_MAX UINT32_MAX UINT64_MAX UINT_FAST8_MAX, }

Figure 2: Example classes, or abstract groupings of symbol types, extracted using document-level information.

Section descriptions Section descriptions are module or class level descriptions.

2 Implementation Details

We report some details about our implementation.

Preprocessing For programming languages that use camel case (e.g., Java), the function representations we normalized representations by replacing camel case boundaries with white space (e.g., myFunction \rightarrow my function). Similarly, languages that use other common word delimiters (e.g., - in the lisp languages) are preprocessing in a similar fashion.

Haskell type declarations are converted into a linear form, or a polish notation, by converting function and container symbols into predicates (e.g., fn :: (a, b) \rightarrow a would be normalized as: fn $tuples_2$ a b a). A similar idea is used in (Andreas et al., 2013). An example

Haskell representation and tree is show in Figure 3.

Pseudo lexicons When training the translation models, we found that improvements can be achieved for latin character language datasets by adding matching component terms to the parallel training data. Such as idea is similar to the use of NP-lists in semantic parsing (e.g., Andreas et al. (2013)), and is a common trick used in other MT tasks (MacCartney et al., 2008).

Learning Parameters Standardly, all hyper parameters in Algorithm 2 are tuned to validation sets (i.e., number of iterations, learning rate, ..). Early stopping is done by monitoring training performance after each iteration using validation sets. Following (Zettlemoyer and Collins, 2009), the full learning rate α in line 7 is defined as $\alpha = \frac{\alpha_0}{1+cso}$, where o = i + t * n, thus making *c* and



Figure 3: An example Haskell linear representation and tree structure.

 α_0 the hyper-parameters.

3 Feature Details

Figure 5-6 show all the feature used. Since our datasets differ in terms of background resources, the green shows the features used in the top models for each dataset. While we did not do rigorous feature testing for this paper, we did notice that the Hiero rules sometimes negatively effect the results, and removed these feature accordingly.

Feature Selection A greedy, backward search selection method is employed for some datasets where overfitting seemed to be an issue. This is done in the following way: after training a complete model, features or templates that lead to incorrect predictions on the validation are greedily removed, and those whose removal increases the accuracy on the validation are shut off. The model is then retrained used the resulting selected set of features or templates. More details are documented in our source code release.

Phrase Features Phrase features are extracted for each input-ouput pair using standard wordbased phrase heuristics over symmetric alignments (i.e., alignments from text \rightarrow components, components \rightarrow text). We use the grow-diag heuristic in all cases, see background in (Koehn, 2009).

Hierarchical phrase rules are extracted using a SAMT style rule extraction procedure (Zollmann and Venugopal, 2006). This can be informally described in the following way: when extracting phrases of the form $X \rightarrow$ english word span || foreign side (using standard methods), sub alignments on each side that match tree patterns are replaced with NTs that match the name of the tree. The LHS of the rule X is also replaced, with with the tree name or the rule that results from combining the inner tree patterns using a small glue grammar. For example:

function \rightarrow this function [function] ||| [function]

Each symbol in brackets is a inner alignmenttree pattern. Given a text and component pair, a CKY-style chart procedure is used for finding all rules subject to the alignment.

Reproducibility 4

We are releasing all data reported on in this paper, as well as all the software used to complete the experiments. Please check the first author's webpage (http://www.ims.unistuttgart.de/institut/mitarbeiter/kyle), and the following Github (https://github.com/yakazimir) for more information.

References

- Jacob Andreas, Andreas Vlachos, and Stephen Clark. 2013. Semantic parsing as machine translation. In in Proceedings of ACL-2013. pages 47-52.
- Huijing Deng and Grzegorz Chrupała. 2014. Semantic approaches to software component retrieval with English queries. In Proceedings of LREC-14. pages 441-450.
- Philipp Koehn. 2009. Statistical Machine Translation. Cambridge University Press.
- Bill MacCartney, Michel Galley, and Christopher D Manning. 2008. A phrase-based alignment model for natural language inference. In Proceedings of EMNLP-2008. pages 802-811.
- Kyle Richardson and Jonas Kuhn. 2014. UnixMan corpus: A resource for language learning in the Unix domain. In Proceedings of LREC-2014.
- Luke S. Zettlemoyer and Michael Collins. 2009. Learning context-dependent mappings from sentences to logical form. In Proceedings of ACL-2009. pages 976-984.
- Andreas Zollmann and Ashish Venugopal. 2006. Syntax augmented machine translation via chart parsing. In Proceedings of the Workshop on Statistical Machine Translation. pages 138–141.

id	descriptions	Java	Ruby	PHP_{en}	Python	Elisp	Haskell	Clojure	С	Unix	Scheme
1	model rank positions								_		
2	english unigrams										
3	foreign unigrams										
4	e/f unigram pairs								-		
5	# unigram matches										
6	# unigram containments										
7	type of unigram matches										
8	# bigram matches	ĺ									
9	# bigram containments										
10	foreign output length										
11	tree position of unigram contain.										
12	tree position bigram matches										
13	viterbi alignment pos.										
14	tree pos. of alignment										
15	phrase instances								_		
16	# known phrases										
17	# matching phrases										
18	# phrase containments										
19	tree position of phrases										
20	tree position matching phrases										
21	tree position phrase contain.										
22	tree position phrase overlap										
23	size of phrase word overlap.										
24	size of english phrase in matched										
25	size of foreign phrase in matched										
26 27	size of english phrases size of foreign phrases										
27	size of english overlapping phrases										
28	size of foreign overlapping phrases										
30	hiero phrase rule										
31	# known hiero rules										
32	# hiero rules with reordering										
33	type of hiero reordering										
34	english sides of hiero rules										
35	foreign side of hiero rules										
36	# unknown hiero rules										
37	unigram pair in description										
38	# of pairs in description										
39	unigram pair in abstract class										
40	unigram in see-also pair								-		
41	unigram in see-also pair match										
42	tree position of item in description										
43	foreign abstract classes seen										
44	type of english unigrams in descriptions										
45	type of foreign words with descriptions										
46	tree position of see-also pair										
47	13+37										
48	13+14+37										
49	5+13+14+37										
50	5+14+37										
51	15+40										
52	31+40										
53	english phrases and abstract classes										
54	english phrases in descriptions										
55	hiero english side and see-also										
56	hiero foreign side and see-also										

Figure 4: Description of features for our English datasets. Green shading shows that a particular feature was used in best model.

1 model rank positions 2 english unigrams 4 eff unigram matches # unigram natches # # unigram natches # # bigram containments 10 foreign output length 11 tree position o fungram contain. 12 tree position o fungram contain. 13 tree position o fungrams contain. 14 tree position of phrases 16 # known phrases 17 # natching phrases 18 # phrase containments 19 tree position phrases contain. 12 tree position phrase contain. 13 tree position phrase contain. 14 tree position phrase contain. 15 tree position phrase contain. 16 # known hiero rules 17 size of foreign phrases 18 # phrase in matched 25 size of foreign overlapping phrases	id	descriptions	PHP_{fr}	PHP_{es}	PHP_{ja}	PHP_{ru}	PHP_{tr}	PHP_{de}
2 english unigrams 3 foreign unigrams 4 eff unigram pairs 5 # unigram natches 7 type of unigram matches 8 # bigram containments 10 foreign output length 11 tree position bigram matches 9 # bigram containments 10 foreign output length 11 tree position bigram matches 12 tree position furgram contain. 13 tree position of unigram contain. 14 tree position phrases 14 tree position phrases 15 # trace instances 16 # known phrases 17 # matching phrases 20 tree position phrase contain. 21 tree position phrase contain. 22 tree position phrase contain. 23 size of phrase word overlap. 33 size of foreign phrases 34 # hiero rules 35 size of oreign phrases 36 size of oreign overlapping phrases 37 size of oreign overlapping phrases<	1	model rank positions						
4 $e^{f'}$ unigram matches5# unigram matches7type of unigram matches9# bigram containments10foreign output length11tree position bigram matches12tree position bigram matches13tree position bigram matches14tree position bigram matches15phrase instances16# known phrases17# matching phrases18# phrase containments19tree position of phrases20tree position of phrases21tree position phrase contain.22tree position phrase contain.23size of foreign phrase contain.24size of forging phrase in matched25size of forging phrase in matched26size of forging phrase in matched27size of forging phrase in matched28size of forging phrases29size of forging phrases20tree position phrase outlapping phrases31# known hiero rules32# known hiero rules33type of hiero rules matched34english sides of hiero rules35foreign abstrat classes seen46piars in description37tree position of item rules38# of pairs in description39tree position of se-also pair41unigram is ex-also pair42tree position of se-also pair43tree position of se-also pair44ther	2	english unigrams						
4 e^{fI} unigram pairs5# unigram matches6# unigram matches7type of unigram matches9# bigram containments10foreign output length11tree position bigram matches12tree position bigram matches13tree position bigram matches14tree position bigram matches15phrase instances16# known phrases17# matching phrases18# phrase containments19tree position of parases20tree position phrase contain.21tree position phrase contain.22tree position phrase contain.23size of english phrase in matched24size of foreign phrase contain.25size of foreign phrase in matched26size of foreign phrase in matched27size of foreign phrase in matched28size of foreign phrases29size of foreign phrases21# known hiero rules22file roules with reordering23type of hiero rules matched24file roules with reordering25size of of oreign overlapping phrases30tip of phrase rule scription31# known hiero rules33file on rules34fole roules pair match35foreign abstrate class36foreign abstrate classes seen37tree position of item in description38# of pairs in descrip	3	foreign unigrams						
6 $\# unique and containments7type of unigram matches9\# bigram matches9\# bigram matches9\# bigram numbers10foreign output length11tree position bigram matches12tree position bigram matches13viterbi alignment pos.14tree position bigram matches15phrase instances16\# known phrases17\# matching phrases18\# phrase containments19tree position of phrase so verlap21tree position phrase in matched23size of of neign phrases in matched24size of foreign ophrases27size of foreign phrase in matched28size of foreign phrases29size of oreiglish phrases20size of oreiglish phrases21size of oreign phrase in matched25size of foreign phrases30hiero phrase rule31type of hiero redering32type of hiero redering33type of hiero redering34# winkown hiero rules47tree position of item rules48foreign abstract class49unigram in se-also pair matched35type of foreign words with description49unigram in se-also pair matched36tree position of item rules47tree position of see-also pair matched38top foreign words with description49unigram $	4							
7type of unigram matches8# bigram matches9# bigram containments10foreign output length11tree position of unigram contain.12tree position of unigram contain.13viterbi alignment pos.14tree pos.15phrase instances16# known phrases17# matching phrases18# phrase containments19tree position of phrases20tree position of phrases21tree position phrase contain.22tree position phrase contain.23size of english phrase in matched25size of foreign phrase in matched26size of foreign phrase matched27size of foreign phrase28size of foreign phrase29size of foreign phrase20tree position phrase phrases21size of of phrase word phrage22size of foreign phrase32size of foreign phrase33size of foreign phrase34size of foreign phrase35size of foreign overlapping phrases36foreign side of hiero rules37tunigram pair in description38foreign abstrat class39unigram pair in description30time position of item in description31foreign abstrat class32tree position of item in description34foreign abstrat class35type of foreign words with descript	5	# unigram matches						
# bigram matches # bigram containments 10 foreign output length 11 tree position of unigram contain. tree position bigram matches viterbi alignment pos. 14 tree pos. of alignment 15 phrase instances 16 # known phrases 17 # matching phrases 18 # phrase containments tree position of phrases tree position matching phrases 20 tree position phrase contain. tree position phrase contain. tree position phrase ord overlap. size of english phrase in matched size of foreign phrase in matched size of foreign phrase in matched size of foreign phrases size of foreign phrases size of foreign phrases size of foreign phrases size of foreign phrases size of of foreign phrases size of foreign phrases size of foreign orlapping phrases size of foreign orlapping phrases size of foreign phrase nutched size of foreign orlapping phrases size of foreign orlapping phrases size of foreign orlapping thrases size of foreign orlapping thrases size of foreign or	6	# unigram containments						
9# bigram containments foreign output length10foreign output length11tree position of bigram matches12tree pos. of alignment phrase instances13phrase instances14tree pos. of alignment phrase containments15matching phrases16# known phrases17# matching phrases18# phrase containments19tree position of phrases contain.10tree position phrase contain.11tree position phrase contain.12tree position phrase contain.13size of english optrase in matched24size of english optrases25size of english optrases26size of english optrases27size of foreign phrases28size of foreign phrases29size of foreign overlapping phrases31# known hiero rules32# hiero rules with reordering33tree position of itero rules34english sides of hiero rules35foreign side of hiero rules36foreign abstract class37unigram in aescription38tree position of set-also pair39tree position of set-also pair31tree position of set-also pair32tree position of set-also pair33tree position of set-also pair34tree position of set-also pair35tree position of set-also pair34tree position of set-also pair <td>7</td> <td>type of unigram matches</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>	7	type of unigram matches						
10 foreign output length 11 tree position of unigram contain. 12 tree position bigram matches 13 tree pos. of alignment pos. 14 tree pos. of alignment phrases 16 # known phrases 17 # matching phrases 18 # phrase containments 19 tree position of phrases 10 tree position phrase contain. 11 tree position phrase overlap 12 size of foreign phrase in matched 13 size of foreign phrase in matched 14 size of foreign phrases 15 size of foreign phrases 16 # known hiero rules 17 size of foreign phrases 18 # known hiero rules 19 tree position phrase rule 11 # known hiero rules 12 # hiero rules 13 # known hiero rules 14 migram pair in abstract class 13 # of pairs in description 14 unigram in se-also pair 14 unigram in se-also pair 14	8	# bigram matches						
11 tree position of unigram contain. 12 tree position bigram matches 13 viterbi alignment pos. 14 tree pos. of alignment 15 phrase instances 16 # known phrases 17 # matching phrases 18 # phrase containments 19 tree position of phrases 10 tree position phrase contain. 11 tree position phrase contain. 12 tree position phrase contain. 11 tree position phrase in matched 12 size of foreign phrases 13 size of foreign phrases 14 tree position nuched 15 size of foreign phrases 16 size of foreign phrases 17 size of foreign phrases 18 size of foreign phrases 19 size of foreign phrases 11 # known hiero rules 12 # hiero rules with reordering 13 type of hiero reordering 14 two position of iter also pair 13 # hiero rules 14 finipa	9	# bigram containments						
12 tree position bigram matches 13 viterbi alignment tos. 14 tree pos. of alignment 15 phrase instances 16 # known phrases 17 # matching phrases 18 # phrase containments 19 tree position of phrases 20 tree position phrase contain. 21 tree position phrase contain. 22 tree position phrase in matched 23 size of english phrase in matched 24 size of foreign phrases 25 size of foreign phrases 26 size of foreign phrases 27 size of foreign phrases 28 size of foreign phrases 29 size of foreign phrases 21 size of foreign phrases 22 size of foreign phrases 33 type of hiero rules 4 english sides of hiero rules 5 foreign side of hiero rules 4 english sides of hiero rules 5 foreign side scription 30 unigram pair in abstract class 41 <	10	foreign output length						
13 viterbi alignment pos. 14 tree pos. of alignment 15 phrase instances 16 # known phrases 17 # matching phrases 18 # phrase containments 19 tree position of phrases 19 tree position phrase contain. 11 tree position phrase overlap. 13 size of phrase word overlap. 14 size of english phrase in matched 15 size of english phrases 16 size of english phrases 17 size of english phrases 18 size of english phrases 19 size of foreign phrases 11 size of english phrases 12 size of foreign phrases 13 size of english overlapping phrases 14 theor outles 17 theor outles 18 known hiero rules 19 thiero rules with reordering 19 english sides of hiero rules 19 unigram in see-also pair 10 unigram in see-also pair 11 unigram in see-also p	11	tree position of unigram contain.						
14tree pos. of alignment15phrase instances16# known phrases17# matching phrases18# phrase containments19tree position of phrases20tree position phrase contain.21tree position phrase overlap.23size of english phrase in matched24size of english phrases25size of english phrases26size of english phrases27size of foreign overlapping phrases28size of foreign overlapping phrases29size of foreign overlapping phrases20size of foreign overlapping phrases27size of foreign overlapping phrases28size of foreign overlapping phrases30hiero phrase rule31# known hiero rules32# hiero rules with reordering33type of hiero reles34# onglish sides of hiero rules35foreign side of hiero rules36# unigram pair in description37unigram in see-also pair41unigram in see-also pair42tree position of item in descriptions44type of foreign words with descriptions45tree position of see-also pair4813+14+37495+13+14+375115+405231+4053hiero english phrases and abtract classes54english phrases and abtract classes55hiero english of alse-calso	12	tree position bigram matches						
15 phrase instances 16 # known phrases 17 # matching phrases 18 # phrase containments 19 tree position of phrases 20 tree position phrase cortain. 21 tree position phrase coverlap 23 size of phrase word overlap. 24 size of foreign phrase in matched 25 size of foreign phrase in matched 26 size of foreign phrases 27 size of english phrases 28 size of foreign phrases 29 size of foreign overlapping phrases 30 hiero rules 31 type of hiero roules 32 # known hiero rules 33 type of hiero roules 34 english sides of hiero rules 35 foreign adde of hiero rules 47 unigram pair in description 38 gize of english unigrams in in adstract class 40 unigram in sec-also pair 41 unigram in sec-also pair 42 tree position of item in descriptions 45 tree position of item in descriptions <td>13</td> <td>viterbi alignment pos.</td> <td></td> <td></td> <td></td> <td></td> <td></td> <td></td>	13	viterbi alignment pos.						
16 # known phrases 17 # matching phrases 18 # phrase containments 19 tree position of phrases 20 tree position phrase contain. 21 tree position phrase contain. 22 size of phrase word overlap. 23 size of foreign phrase in matched 26 size of foreign phrases 27 size of foreign phrases 28 size of foreign phrases 29 size of foreign phrases 29 size of foreign phrases 29 size of foreign overlapping phrases 30 type of hiero rules 4 inter oules with reordering 31 type of hiero rules 36 foreign side of hiero rules 37 unigram pair in description 38 # of pairs in description 39 unigram pair in abstract classes 40 unigram is ea-also pair 41 unigram is in descriptions 45 type of foreign words with descriptions 46 type of foreign solo pair 41 type of foreign words with descriptio	14	tree pos. of alignment						
17 # matching phrases 18 # phrase containments 19 tree position of phrases 20 tree position phrase contain. 21 tree position phrase coverlap 23 size of phrase word overlap. 24 size of foreign phrase in matched 25 size of foreign phrases 28 size of english phrases 29 size of foreign phrases 28 size of foreign phrases 29 size of foreign phrases 29 size of foreign phrases 20 thero phrase rule 31 # known hiero rules 32 # hiero rules with reordering 33 type of hiero rules 34 english sides of hiero rules 35 foreign side of hiero rules 36 # unknown hiero rules 37 unigram pair in abstract class 38 # of pairs in description 39 unigram in see-also pair 41 unigram in see also pair 42 tree position of see-also pair 43 type of foreign words with descriptions	15	phrase instances						
18 # phrase containments 19 tree position of phrases 20 tree position phrase contain. 21 tree position phrase contain. 22 tree position phrase contain. 23 size of phrase word overlap. 24 size of oreign phrase in matched 25 size of oreign phrases 26 size of oreign phrases 27 size of oreign phrases 28 size of foreign overlapping phrases 29 size of foreign overlapping phrases 30 thero rules with reordering 31 # hiero rules 34 english sides of hiero rules 35 foreign side of hiero rules 36 # unknown hiero rules 37 unigram pair in description 38 # of pairs in description 39 unigram in see-also pair 41 unigram in see-also pair 42 type of english unigrams in descriptions 44 type of english unigrams in descriptions 45 tree position of see-also pair 44 type of english unigrams in descriptions	16	# known phrases						
19tree position of phrases20tree position phrase contain.21tree position phrase overlap23size of phrase word overlap.24size of english phrase in matched25size of foreign phrase in matched26size of foreign phrase in matched27size of foreign phrases28size of foreign phrases29size of foreign overlapping phrases20hiero phrase rule31# known hiero rules31# known hiero rules33type of hiero rordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram pair in see-also pair41unigram pair si description43foreign abstract class seen44type of foreign words with descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37505+13+13+14+375115+4053english phrases and abstract classes54english phrases and abstract classes55hiero english side and see-also								
20tree position matching phrases21tree position phrase contain.22tree position phrase overlap33size of phrase word overlap.44size of english phrase in matched25size of foreign phrases26size of english phrases27size of oreign phrases28size of foreign phrases29size of foreign overlapping phrases30hiero phrase rule31# known hiero rules34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair42type of foreign words with descriptions43foreign words with descriptions44type of foreign words with descriptions45type of foreign words with descriptions46tree position of see-also pair4713+37495+13+14+37505+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases and abstract classes54english phrases and abstract classes54english phrases and abstract classes55hiero english side and see-also								
21tree position phrase contain.22tree position phrase overlap.23size of phrase word overlap.24size of english phrase in matched25size of foreign phrase in matched26size of english phrases27size of foreign phrase28size of english overlapping phrases29size of foreign overlapping phrases20hiero phrase rule31# known hiero rules32# hiero rules with reordering33type of hiero rules34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair42tree position of item in descriptions43foreign abstract classe seen44type of foreign words with descriptions45type of foreign words with descriptions46tree position of see-also pair4713-374813-14+37505+13+14+375115+405315+4054english phrases and abstract classes54english phrases and abstract classes54english phrase and abstract classes54english phrases and abstract classes55hiero english ide and see-also								
22 tree position phrase overlap 3ize of pirase word overlap. 24 size of english phrase in matched 25 size of foreign phrases 27 size of english phrases 28 size of english overlapping phrases 29 size of foreign overlapping phrases 20 size of foreign overlapping phrases 29 size of foreign overlapping phrases 30 hiero phrase rule 31 # known hiero rules 32 # hiero rules with reordering 33 type of hiero rules 36 # unknown hiero rules 37 unigram pair in description 38 # of pairs in description 39 unigram in see-also pair match 41 unigram in see-also pair match 42 tree position of item in description 44 type of english unigrams in description 45 type of foreign words with descriptions 46 tree position of see-also pair 47 13+137 48 13+14+37 49 5+13+14+37 50 5+14+437 <t< td=""><td></td><td></td><td></td><td></td><td></td><td></td><td></td><td></td></t<>								
23 size of phrase word overlap. 24 size of english phrase in matched 25 size of foreign phrases 28 size of english phrases 29 size of foreign phrases 29 size of foreign overlapping phrases 30 hiero phrase rule 31 # known hiero rules 24 english of hiero rules 35 foreign side of hiero rules 36 # unknown hiero rules 37 unigram pair in description 38 # of pairs in description 39 unigram in sec-also pair 41 unigram in sec-also pair 41 unigram in sec-also pair 42 tree position of item in description 43 foreign words with descriptions 44 type of foreign words with descriptions 45 type of foreign words with descriptions 46 tree position of sec-also pair 47 13+37 48 13+14+37 49 5+13+14+37 50 5+13+14+37 51 15+40 52 31+40								
24 size of english phrase in matched 25 size of foreign phrases in matched 26 size of foreign phrases 27 size of foreign phrases 28 size of foreign overlapping phrases 29 size of foreign overlapping phrases 30 hiero phrase rule 31 # known hiero rules 32 # hiero rules with reordering 33 type of hiero rules 36 # unknown hiero rules 37 unigram pair in description 38 # of pairs in description 39 unigram in sec-also pair 41 unigram in sec-also pair 41 unigram in description 43 foreign abstract class 44 type of foreign words with descriptions 45 type of foreign words with descriptions 46 type of foreign words with descriptions 47 13+37 48 13+14+37 49 5+13+14+37 50 5+14+37 51 lsf-40 52 31+40 53 inero english phrases and abs								
25size of foreign phrase in matched26size of english phrases27size of foreign phrases28size of foreign overlapping phrases29size of foreign overlapping phrases30hiero phrase rule31# known hiero rules24# hiero rules with reordering33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram pair in abstract class41unigram pair in description43foreign abstract class seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4113+374813+14+37495+13+14437505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also	-							
26size of english phrases27size of foreign phrases28size of foreign portagping phrases29size of foreign overlapping phrases30hiero phrase rule31# known hiero rules32# hiero rules with reordering33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram in see-also pair41unigram in see-also pair41unigram in see-also pair41type of foreign words with descriptions45type of foreign words with descriptions46type of foreign words with descriptions4713+374813+14+37495+13+14+37505+14+4375115+405231+4053english phrases and abstract classes54english pirases in descriptions55hiero english side and see-also								
27size of foreign phrases28size of english overlapping phrases29size of foreign overlapping phrases30hiero phrase rule31# known hiero rules32# hiero roles with reordering33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair42tree position of item in descriptions43foreign abstract classes seen44type of foreign words with descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+375115+405231+4053english phrases and abstract classes55hiero english side and see-also	-							
28size of english overlapping phrases29size of foreign overlapping phrases30hiero phrase rule31# known hiero rules22# hiero rules with reordering33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram in see-also pair41unigram in see-also pair match42tree position of item in descriptions43foreign words with descriptions44type of oreign words with descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37505+14+375115+405231+4053english phrases and abstract classes54english side and see-also								
 size of foreign overlapping phrases hiero phrase rule # known hiero rules # hiero rules with reordering type of hiero rordering type of hiero rordering english sides of hiero rules foreign side of hiero rules # of pairs in description # of pairs in description unigram pair in abstract class unigram in see-also pair match tree position of item in descriptions foreign abstract classes seen tree position of see-also pair tree position abstract classes tree plish phrases and abstract classes english phrases in descriptions hiero english side and see-also 								
30hiero phrase rule31# known hiero rules32# hiero rules with reordering33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
31# known hiero rules32# hiero rules with reordering33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram in sec-also pair41unigram in sec-also pair41unigram in sec-also pair41tree position of item in descriptions43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of sec-also pair4713+374813+14+37495+13+14+37505+13+1375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and sec-also								
32# hiero rules with reordering33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair match42tree position of item in descriptions43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
33type of hiero reordering34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair match42tree position of item in descriptions43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also	-							
34english sides of hiero rules35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair match42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
35foreign side of hiero rules36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair match42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+37505+14+375115+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
36# unknown hiero rules37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+37505+14+1375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also	-							
37unigram pair in description38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair match42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
38# of pairs in description39unigram pair in abstract class40unigram in see-also pair41unigram in see-also pair match42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
 unigram pair in abstract class unigram in see-also pair unigram in see-also pair match tree position of item in description foreign abstract classes seen type of english unigrams in descriptions type of foreign words with descriptions tree position of see-also pair tree position of see-also pair ta+37 13+37 13+14+37 51 15+40 53 english phrases and abstract classes english phrases in descriptions there english side and see-also 								
40unigram in see-also pair41unigram in see-also pair match42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
41unigram in see-also pair match42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
42tree position of item in description43foreign abstract classes seen44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+1374813+14+374813+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
 43 foreign abstract classes seen 44 type of english unigrams in descriptions 45 type of foreign words with descriptions 46 tree position of see-also pair 47 13+37 48 13+14+37 49 5+13+14+37 50 5+14+37 51 15+40 52 31+40 53 english phrases and abstract classes 54 english phrases in descriptions 55 hiero english side and see-also 								
44type of english unigrams in descriptions45type of foreign words with descriptions46tree position of see-also pair4713+374813+14+37495+13+14+37505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
 45 type of foreign words with descriptions 46 tree position of see-also pair 47 13+37 48 13+14+37 49 5+13+14+37 50 5+13+14+37 50 5+14+37 51 15+40 52 31+40 53 english phrases and abstract classes 54 english phrases in descriptions 55 hiero english side and see-also 	-							
46 tree position of see-also pair 47 13-37 48 13+14+37 49 5+13+14+37 50 5+13+14+37 51 15+40 52 31+40 53 english phrases and abstract classes 54 english phrases in descriptions 55 hiero english side and see-also								
47 13+37 48 13+14+37 49 5+13+14+37 50 5+14+37 51 15+40 52 31+40 53 english phrases and abstract classes 54 english phrases in descriptions 55 hiero english side and see-also								
48 13+14+37 49 5+13+14+37 50 5+14+37 51 15+40 52 31+40 53 english phrases and abstract classes 54 english phrases in descriptions 55 hiero english side and see-also								
49 5+13+14+37 50 5+14+37 51 15+40 52 31+40 53 english phrases and abstract classes 54 english phrases in descriptions 55 hiero english side and see-also								
505+14+375115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also	-							
5115+405231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
5231+4053english phrases and abstract classes54english phrases in descriptions55hiero english side and see-also								
 53 english phrases and abstract classes 54 english phrases in descriptions 55 hiero english side and see-also 								
54 english phrases in descriptions 55 hiero english side and see-also	-							
55 hiero english side and see-also								
interesting in side and see also								
	1 50 1	mere rereign side und see uise						

Figure 5: Feature information for our non-English datasets.