ALIGNING THE RTE 2006 CORPUS

MICROSOFT RESEARCH TECHNICAL REPORT MSR-TR-2007-77

Chris Brockett
Natural Language Processing Group
Microsoft Research
chris.brockett@microsoft.com

INTRODUCTION

The past several years have seen much growth of research interest in determining when sentences or phrases "mean the same thing," for purposes ranging from information retrieval to summarization. One consistent obstacle to research in this area has been the lack of shared annotated corpora with which to measure and compare the effectiveness of algorithms and applications for determining and measuring semantic similarity. The PASCAL Recognizing Textual Entailment challenges (Dagan et al. 2006; Bar-Haim et al. 2006) have made significant progress in remedying this gap by making available to the research community collections of sentence pairs that have been annotated to indicate whether an inference can be drawn between them.

While lexical and phrasal similarity are far from the only features of use in drawing inferences (see Vanderwende et al., 2006 for discussion), they present fertile ground for exploration, and the RTE data is potentially applicable for purposes beyond those for which it may have been initially intended, for example, in measuring or calibrating synonym extraction and paraphrase identification. In the past, the Natural Language Processing Group at Microsoft Research has made available to the research community a corpus of several thousand paraphrase sentence pairs (Dolan and Brockett, 2005); this corpus, however, has had the drawback that it was extracted by largely automatic techniques that limit its effectiveness in evaluating some tasks. For this reason, we have decided to release to the research community versions of the 2006 PASCAL RTE development and test corpora in which semantically equivalent words and phrases in the Text and Hypothesis sentences are aligned in a manner analogous to the alignments in statistical machine translation. In the hope that these resources will assist in evaluating semantic similarity word- and phrase-aligned datasets we have made them available via the Textual Entailment Resource Pool at the following location:

http://aclweb.org/aclwiki/index.php?title=Textual_Entailment_Resource_Pool

DATA FORMAT

The annotated files are collected in two data sets corresponding to the RTE 2006 development and test sets. Each data set contains three files in utf8 format, one for each of three annotators.

Each file contains a list of text-hypothesis pairs in the following format:

- A line marking the beginning of the text-hypothesis pair,
- A line containing the text sentence, tokenized Penn Treebank style.
- A line containing the hypothesis sentence, tokenized Penn Treebank style, each word being followed by a list of indices pointing to the corresponding word or words in the text sentence.

Indices prefixed by the letter *p* are POSSIBLE links, the others are SURE. Where no link is specified, i.e., where the word has a no link to the text, the list has been left blank. Indices are 1-based, 0 being theoretically reserved for mappings to NULL in the tool. [1] A sample is given below.

sentence pair 1
ECB spokeswoman , Regina Schueller , declined to comment on a report in Italy 's La Repubblica newspaper that the ECB council will discuss Mr. Fazio 's role in the takeover fight at its Sept. 15 meeting .
NULL ({ / / }) Regina ({ 4 p1 p2 / / }) Shueller ({ 5 p1 p2 / / }) works ({ / / }) for ({ / / }) Italy ({ 14 / / }) 's ({ 15 / / }) La ({ 16 / / }) Repubblica ({ 17 / / }) newspaper ({ 18 / / }) . ({ 38 / / })

CITATION

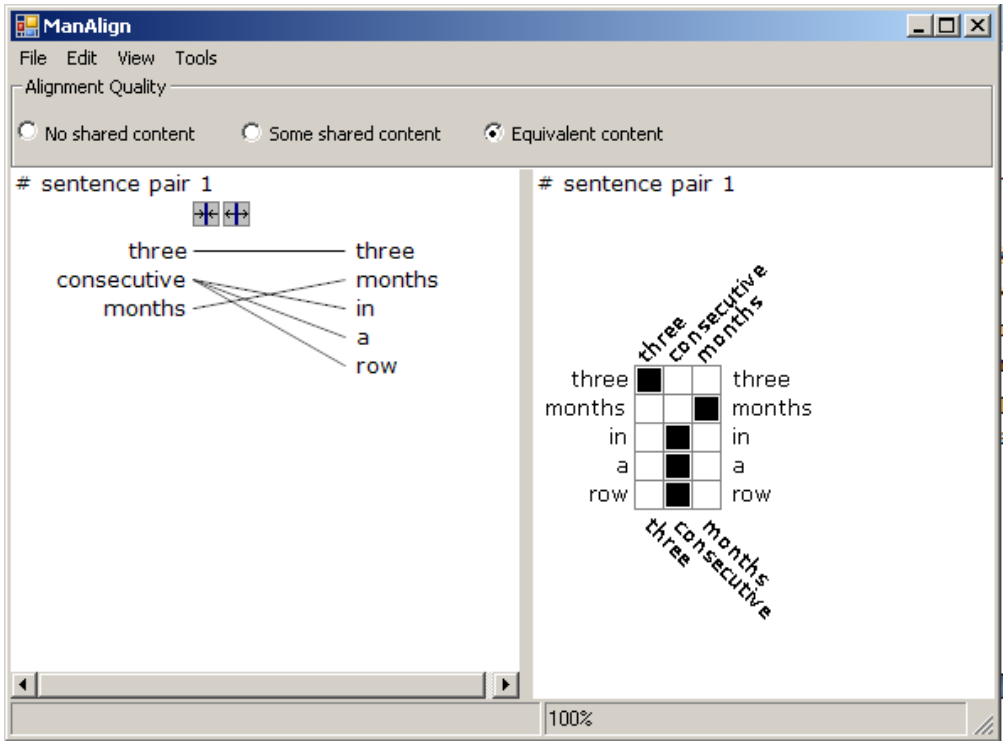
If you use these aligned datasets in your research, we ask that you cite this technical report.

[1] This is one of several features of the output of the annotation tool that can be safely ignored in these datasets. The word NULL at the beginning of the hypothesis string is an artifact of using the annotation tool to view bidirectional machine translation alignments. The spaces following slashes in the tags, e.g., ({ 18 / / }) are unused in this task.

THE ANNOTATION PROCESS

Three annotators were contracted to undertake the word and phrase alignments. Their identities are obfuscated in the datasets by the letters A, B and C (maintained across both development and test sets.) The annotators were asked to insert links between words and phrases of similar meaning in the two sentences in accordance with a short set of guidelines (see APPENDIX: GUIDELINES FOR ANNOTATORS for details), developed on the basis of initial trials aligning the RTE 2005 corpus. The guidelines were not intended to be definitive, but to provide a framework around which the annotators might develop intuitions about how to go about tagging. The development set was treated as training data, and the annotators were actively encouraged to discuss issues arising out of this data both among themselves and with the present author, using email to raise and resolve questions and reconcile results. In processing the test set, however, they were instructed not to discuss the annotations either among themselves or with the author.

Following the methodology employed by Och and Ney (2000, 2003) in evaluating statistical machine translation alignments, the annotators were instructed to identify positive links as either SURE or POSSIBLE. The process was facilitated using a GUI tool originally developed for manually aligning words for evaluating multilingual machine translation systems. This permitted the annotators to view sentence pairs both as a matrix and as a pair of parallel strings of words with lines of association between them. Within the matrix, annotators were able to click on intersecting squares to indicate the degree of confidence they had in the alignment.



In order to avoid biasing their alignment judgments, the annotators were not toldwhich of the RTE Text and Hypothesis sentences contained valid inferences according to the official evaluations,[2] although they were informed that sentences could be dissimilar or unrelated. During training on the development set, annotators reported that they experienced the greatest difficulties with sentence pairs that did not represent valid inferences, since links between words did not form identifiable patterns. In such unrelated pairs, meaningful mappings often proved impossible even at the lexical or phrasal level. In these cases, if a word appeared, for example, multiple times in one sentence and only once in the other, the annotators were advised to align from the left, unless there was a good semantic basis to do otherwise. Similarly, they were advised to ignore function words and focus on content words if the mappings seemed too random or sparse to be meaningful.

ANNOTATOR AGREEMENT

Annotator agreement is difficult to measure meaningfully in these data sets, since the large number of words that map to null (~31% in the development data and ~27% in test data) tends to skew the results.[3] In order to avoid proliferating null links, only mappings from tokens in the (generally briefer) hypothesis sentence to the text sentence were considered in computing annotator agreement. Fleiss Kappa scores were 0.73364 on the development data and 0.72887 on the test data, which can be taken to mean "substantial agreement." For the purposes of computing Kappa scores, we computed agreement on both the link indices and their identity (SURE, POSSIBLE, null).

More concretely, we found that all three annotators concurred on ~70% of proposed links in both data sets, including null links, while two out of three agreed on about ~30% of cases. This can be seen in Table 1, where the agreement patterns are reliably consistent across data sets. Three-way disagreements were exceptionally rare and may reflect indecisiveness or error on the part of one of the annotators.

Data Set	Cumulative Total	3 of 3 Agree		2 of 3 Agree		No Agreement	
		Count	(%)	Count	(%)	Count	(%)
Development	11438	8089	70.72	3326	29.08	23	0.20
Test	13163	9184	69.77	3946	29.98	33	0.25

Table 1. Annotator agreement. (Cumulative totals are the union of all links by all annotators, including null links.)

[2] The annotators were instructed not to check the buttons relating to shared content.

[3] Word alignments in the monolingual RTE data are generally much sparser than those typically found in multilingual alignment, even when a valid inference can be drawn.

Table 2 below shows the counts of SURE, POSSIBLE and null links assigned by each annotator. In general, the annotators seem to have been in close agreement with respect to those links they deemed SURE. (Automatic linking of identical words may have helped consistency here.) The widest variations manifest themselves in the POSSIBLE links. Annotator C appears to have been moderately aggressive in assigning much higher than average POSSIBLE mappings on the development set, while annotator A has been significantly more conservative in creating POSSIBLE links in the test set, resulting in significantly fewer overall links. These variations are further reflected in Table 3 in C's high link fertility in the development set and A's relatively low link fertility in the test set respectively. Table 3 also reveals a higher rate of non-null links assigned to hypothesis sentence words in the test data.

Data Set	Anno-tator	SURE		POSSIBLE		Null		Total Annotations
		Count	%	Count	%	Count	%	
Development	A	5983	54.16	2154	19.50	2909	26.34	11046
	B	6022	54.75	2015	18.32	2963	26.94	11000
	C	5888	50.86	2733	23.61	2956	25.53	11577
Test	A	5690	60.16	1564	16.54	2204	23.30	9458
	B	5719	56.00	2257	22.10	2237	21.90	10213
	C	5728	56.43	2197	21.65	2225	21.92	10150

Table 2: SURE, POSSIBLE and null annotations

Data Set	Annotator	Tokens in Hypothesis Sentences	Tokens with Non-Null Link(s)		SURE + POSSIBLE	Mean Fertility
			Count	%		
Development	A	9429	6520	69.15	8137	1.25
	B	9429	6466	68.58	8037	1.24
	C	9429	6473	68.65	8621	1.33
	Mean		6486	68.79	8265	1.27
Test	A	8325	6121	73.53	7254	1.19
	B	8325	6088	73.13	7976	1.31
	C	8325	6100	73.27	7925	1.30
	Mean		6103	73.31	7718	1.27

Table 3: Non-null token links, with fertilities

ACKNOWLEDGEMENTS

We would like to thank the Butler Hill Group, especially Adam Savel, Annika Hämäläinen and Amy Muia, for their assistance in annotating the corpus.

REFERENCES

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The PASCAL Recognising Textual Entailment Challenge. In Quiñonero-Candela et al., editors, *MLCW 2005*, *LNAI Volume 3944*, pages 177-190. Springer-Verlag.

William B. Dolan and Chris Brockett. 2005. Automatically Constructing a Corpus of Sentential Paraphrases. *IWP2005: The Third International Workshop on Paraphrasing*, Cheju, Korea.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini and Idan Szpektor. 2006. The Second PASCAL Recognising Textual Entailment Challenge. In *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy.

Franz Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of the ACL*: 440-447. Hong Kong, China.

Franz Och and Hermann Ney. 2003. A Systematic Comparison of Various Statistical Alignment Models. *Computational Linguistics* 29(1): 19-52.

Lucy Vanderwende, Deborah Coughlin, Bill Dolan. 2006. What Syntax can Contribute in Entailment Task. In Dagan et. al., 2006.

APPENDIX: GUIDELINES FOR ANNOTATORS

GENERAL

There are two datasets to be annotated. Each comprises 800 pairs of sentences. In half of these pairs, the second sentence may be inferred from the first, and in the other half no inference may be drawn. You will not be told which is which. The first sentence is always the text and the second the hypothesis.

Attempt to align as many words as POSSIBLE in all sentence pairs, even if you think that an inference cannot or should not be drawn between the two sentences. In many cases, alignment may not be POSSIBLE. Expect to find the word alignments in unrelated sentences to be sparse, or to crisscross with no obvious patterns.

Punctuation should be aligned where relevant and consistent with the other alignments.
Don't be surprised to see some differences in quality of the English between the sentences in some pairs, since some sentences may have been mechanically translated.

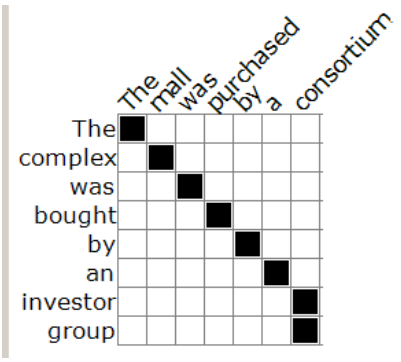
SURE VERSUS POSSIBLE LINKS

The alignment tool supports two degrees of confidence in alignment: SURE and POSSIBLE. In the tool, SURE is indicated by clicking in the matrix to create a black box. To flag a link as POSSIBLE click to create a white box.

SURE LINKS

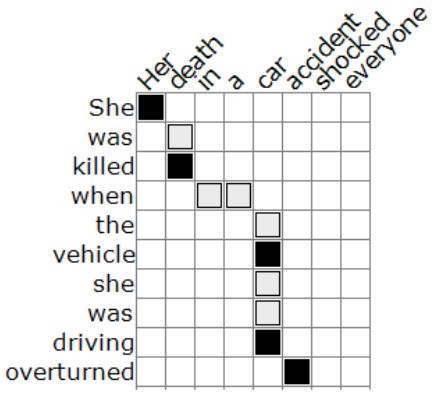
Align as SURE those cases where you are reasonably confident that the words should be mapped. Some reasons why the words should be mapped as SURE are:

- The words are identical. (The tool allows you to automatically link identical words.)
- The words are synonyms or near synonyms.
- One word is a hypernym or hyponym of the other. For example, "mall" and "complex" in the diagram below.

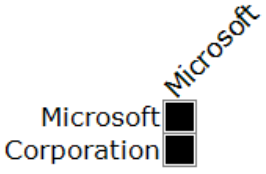


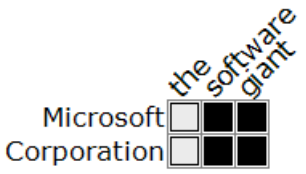
- A word in one sentence is a content word and forms part of a sequence that is synonymous with a content word or sequence of content words in the other sentence (e.g., "investor" in "investor group" should be mapped as SURE to "consortium" in the example above.)

- The words are not synonyms, but are morphologically-derived (e.g., "development" and "develop") or otherwise semantically closely related (e.g., "death" and "kill", "car" and "driving", below. Note that "driving" is actually part of the embedded relative clause.)

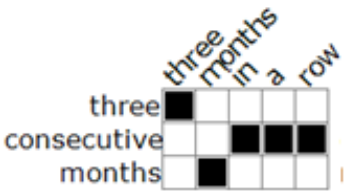


- The words are part of a proper name that has an equivalent: For example "Microsoft Corporation" and "Microsoft". The same also holds for epithets and other descriptive elements when these map to proper names.



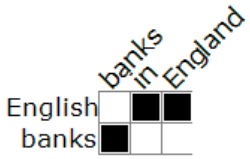


- The words are part of a tightly constructed idiom that maps to another word or phrase: For example, “in a row” and “consecutive” should have SURE mappings:

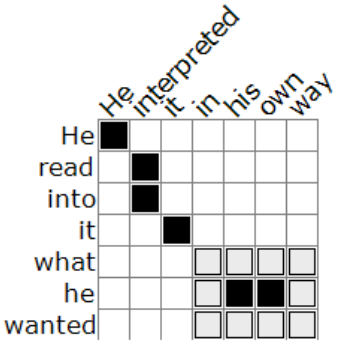


You will probably find that function words generally tend to align as POSSIBLE. This is not always the case, however, and you will need to take into account the relative semantic unity (“tightness”) of a phrase in making the judgment.

- A word is a function word, but in conjunction with an adjacent word matches an adjective, for example, “in England” corresponds to “English” and therefore is assigned SURE.



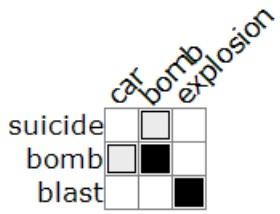
- A word is a function word, but is part of a phrasal verb or is part of the argument structure of the verb: For example, “read into” and “interpret” should have SURE links in the following:



POSSIBLE LINKS

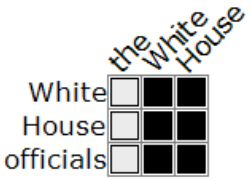
Align as POSSIBLE those cases where you think that it is POSSIBLE that the words might reasonably be linked, but for which you are less confident. Some reasons why the words should be mapped as POSSIBLE rather than SURE are:

- One word is part of a pair of words, and is commonly associated with the word but is not essential to the mapping between meaningful elements. Below, we want to tag "car bomb" and "suicide bomb" as likely synonyms, without asserting that "car and "suicide" are synonymous: the solution is to map "car" and "suicide" to "bomb" as POSSIBLE. (Note additionally that "car" and "suicide" are not linked to each other.)



  The distinction between this case and the case of "investor group" and "consortium" seen earlier is fairly subtle, and will call for some judgment on your part. One could not contextually substitute "car" for "bomb" in quite the same way that one could substitute, say, "investors" for "consortium."

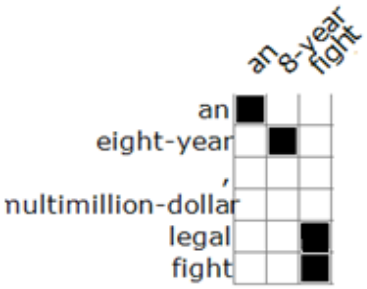
- One of the words is a function word that are required as part of a mapping for structural reasons, but which is not independently semantically equivalent. For example, "the" in the diagram below is mapped as a POSSIBLE link to every other item in the matching phrase.



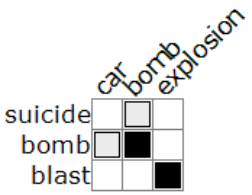
WHEN NOT TO ASSIGN A LINK

Don't assign a link if the semantic association between the words is too distant or not motivated by structure. This is especially true for content words. In the example above, "multimillion-dollar" has not been aligned with "fight" because there is no intrinsic

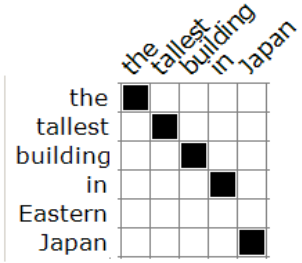
semantic connection, while ”legal” has been aligned as POSSIBLE because it is a more well-defined or definable category of “fight”. Expressions involving numbers may tend to fall into this class of unassignable items.



Don’t attempt to coerce mappings where they are invalid. For example, there is no link between “car” and “suicide” in the following (previously seen) match, so the square in the matrix is left blank, even though these two words are both mapped as having POSSIBLE links to “bomb”



Don’t assign a link if a word significantly changes the meaning of the phrases or would create a relationship between the sentences that does not exist. For example, given the two phrases “the tallest building in Japan” and “the tallest building in eastern Japan”, “eastern” should not be linked to Japan:

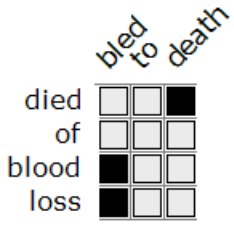


PREFER PARSIMONY

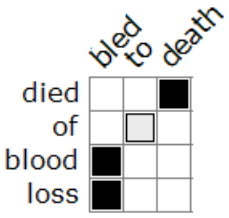
In general, you should try to be reasonably parsimonious in the links you assign. Large blocks of crosshatched links where every word is connected to the other are probably not good alignments. Where POSSIBLE, look for smaller phrases or chunks that you can align more locally.

If two phrases correspond broadly as paraphrases of each other, but there are clearly identifiable matching subphrases, flag the elements in highest confidence subphrase mappings as SURE, and the ones about which you might be less confident as POSSIBLE.

In the following example, the idiom "bleed to death" could potentially map to its counterpart in multiple ways, but an alignment that maps only the subcomponents is to be preferred, since the parts of the overall mapping are clearly separable.

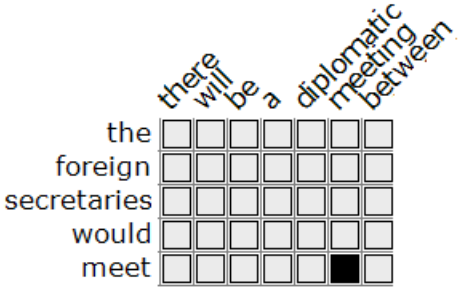


Poor Alignment: Too many words are mapped to each other.

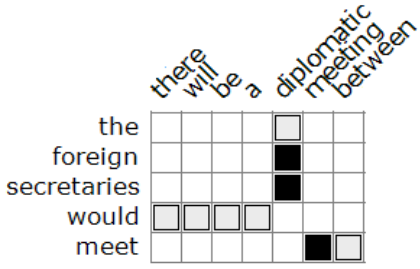


Good Alignment: Only those words that match up are linked semantically or structurally are linked.

Here are some more examples:

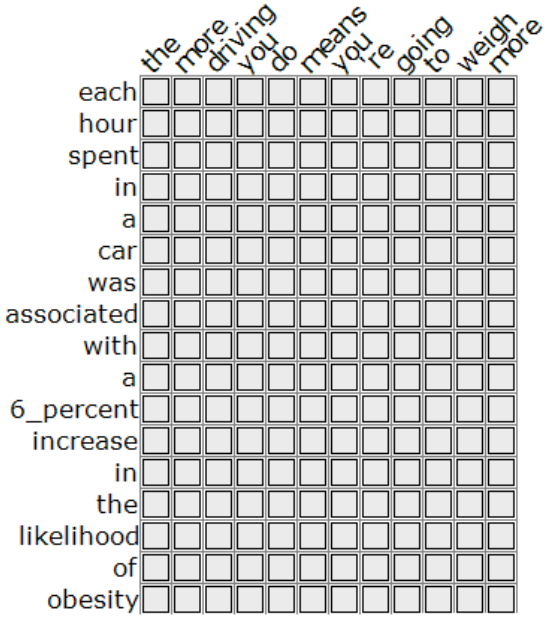


Poor Alignment: There are too many undifferentiated mappings, and the strong semantic relationship between "foreign secretaries" and "diplomatic" is obscured.

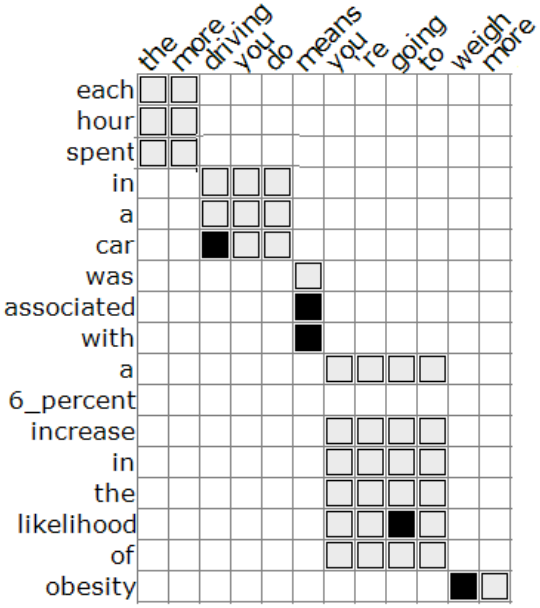


Good Alignment: There are relatively few mappings, containing three learnable chunks. Content words with strong semantic associations are linked as SURE, while function words are associated with the nearest terms as POSSIBLE.

In some cases it may prove impossible to break the mappings up into manageable units. In these cases, flag as SURE any high probability mappings if any, and POSSIBLE the remainder. This can be seen in bottom right corner of the following example.

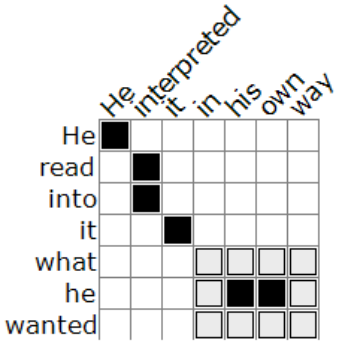


Poor alignment: There are far too many associations to be informative



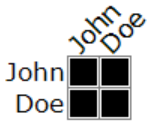
Good alignment: Links are parsimonious, and grouped into phrases, with strongly associated content words linked as SURE.

Another example has been seen earlier in the mapping between "what he wanted" and "in his own way." The diagram is repeated below. Note that the higher probability mappings within the relevant block of links are flagged as SURE.

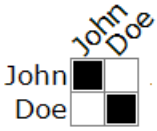


PREFER LINEARITY

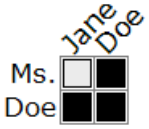
If there is a reasonable choice between a cross-hatched mapping and a one in which the mappings progress in a linear manner, you should prefer the latter. This may pose problems in the case of proper names, where crosslinking may be necessary. In general, however, if you can maintain a one-one mapping you should do so.



Wrong: Every word is mapped to every other word, even though they can be aligned separately.



Right: Individual words are aligned in sequence.



Right: Words are crossmapped because a simple one-one relationship cannot be established. "Jane" and "Ms." are linked as POSSIBLE