Comparing the House and Senate Versions of the National Defense Authorization Act

Laura Odell, Katharine Burton, and Miranda Seitz-McLeese

The Challenge: The current process for tracking changes between the House and Senate versions of a National Defense Authorization Act is manual, repetitive, and time-consuming, leaving little time for analysis.

To save resources and allocate more time to analysis, the Office of Legislative Affairs asked IDA to see whether the IDATA capability could automate the text comparison process.

Background

The DoD Office of Legislative Affairs spends significant resources comparing different versions of the National Defense Authorization Act (NDAA) in a process that has not changed in two decades. The current process for tracking changes between the House and Senate versions of an NDAA is manual and repetitive. Analysts compile tables with the House language on one side and the Senate language on the other, and then examine the text for differences. Simply searching for differences in the text takes up so much time that the subsequent work of analyzing the potential impact of the discovered differences or determining which version is likely to be present in the final version become secondary priorities.

To save resources and allocate more time to analysis, the Office of Legislative Affairs asked IDA to see whether the IDATA capability could automate the text comparison process.

Methodology

Draft legislation is available online in XML format. IDA researchers downloaded the XML files and used the XML structure to split them into smaller sections. Researchers used these smaller sections to generate a term frequency-inverse document frequency (TF-IDF)[1] matrix, and then used latent semantic analysis (LSA)[2] to transform the matrix into a smaller dimensional vector space. Once the points were

[1] TF-IDF weights a given term to determine how well the term describes an individual document within a corpus of documents. It weights a term positively for the number of times the term occurs within a specific document and weights the same term negatively relative to the number of documents that contain it (tfidf.com, http://www.tfidf.com/Accessed September 26, 2017).

[2] LSA is a method for determining the similarity in meaning of words and phrases by analyzing a large corpus of text and producing a set of related concepts and terms. LSA is known to combat the effects of synonymy (a state in which a word is a synonym for other words) and polysemy (that a word or phrase may have more than one meaning).

embedded in this space, the team paired off the points, one from the House version and one from the Senate version per pair, starting with the pair that was closest together according to Euclidean distance. At a certain distance, we considered the points too far away from each other to have a relationship. These unpaired points were labeled "no match."

We then used the point pairings (and the unpaired points) to automatically generate a table. The table was color-coded on a red-yellow-green spectrum, with red indicating a low level of similarity between points, yellow indicating a medium level of similarity, and green indicating a high level of similarity. Figure 1 shows an excerpt of the table, which had 2,982 rows.

Results

The resulting spreadsheets required a human analyst to clean and verify the data. The algorithm sometimes missed connections that it should have made or made unwarranted connections. Overall, however, the algorithm was able, with a high statistical probability, to correctly find sections that were substantially the same. This allowed analysts to concentrate their efforts on the differences between sections.

The algorithm provides substantial time and cost savings for both the analysts and DoD. Because verification is faster than production, analysts require less time to verify or correct an algorithmically produced alignment than to find the same alignment manually. The algorithm

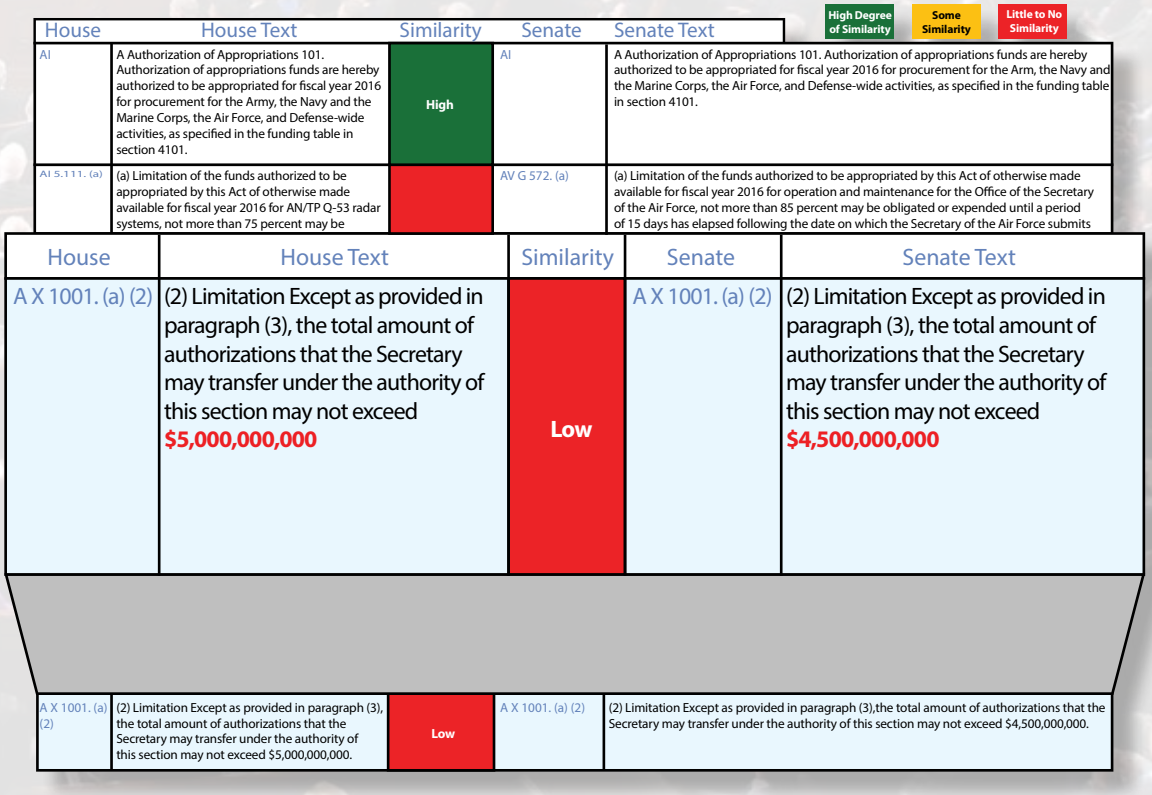

Figure 1. House vs. Senate Language, NDAA 2016

reduces the analysts' role in the initial search for differences – a shift from search and filter to verification and correction.

The time saved can enable the Office of Legislative Affairs to increase throughput without hiring new employees and will allow current employees to focus on tasks that require critical thinking.

References

Black, P.E. 2004. "Euclidean Distance." In *Dictionary of Algorithms and Data Structures,* edited by V. Pieterse and P.E. Black. December 17, 2004. Available from https://xlinux.nist.gov/dads/HTML/euclidndstnc.html. Accessed October 2, 2017.

tfidf.com. http://www.tfidf.com/. Accessed September 26, 2017.