Identify, analyze and synthesize modern efforts into language modeling, distributed (continuous) representations of language (e.g., word, sentence and document embeddings), and/or the application of either (or both) of these two areas. You may choose any subset of the topics.

Your paper may branch into other areas, as long as it includes a subset of the topics. E.g., problems formulated in the noisy-channel model (e.g., machine translation, speech recognition) or as reranking problems (e.g., parsing, machine translation) will use language models. Though language models typically model the surface form of (sequences of) words, researchers will also talk about “X” language models as a way to more explicitly (or better) incorporate “X” into the model formulation and learning. For instance, class-based language models partition words into different classes, or strata, in order to better differentiate aspects of word use, semantic language models are posited to better understand semantics or pragmatics, and entity language models incorporate entity coreference information as explicit features or aspects.

The topics for this paper often work in tandem. For instance, language modeling can give rise to new distributed representations, distributed representations can be used in classification tasks, and lexical features and language modeling scores can be useful parts of classification.

Identify For this assignment you will need to find an appropriate number of papers to discuss in detail. Though the final number that you select is highly dependent on, among other things, which topics you choose, the length of the papers, and their venues, a reasonable number of papers is between five and ten. This range does not constitute required minimums or maximums.

You may read many more papers than you discuss in detail. Do not view this as “wasted” effort—these should help inform the overall narrative and context for your discussion.

Analyze Ask and answer fundamental research questions: what were the goals of each of the papers? What scientific and engineering questions did each of the tackle? How well did the evaluations support the main claims? What was not done that could have been done?

Synthesize How do the efforts relate to one another? Do they follow one after another, making (incremental) progress on a task (metric)? Does one question some basic assumptions of another, and if so, how do the other papers fit in? What are the limitations of these approaches, and what still remains to be done? You can also link these papers and ideas to related fields.
Requirements Papers should be four pages, not including references, in the ACL format. Please use the ACL 2017 style guide; both \TeX{} and Microsoft Word (docx) versions are on GL:

\url{/afs/umbc.edu/users/f/e/ferraro/pub/sty/acl17-l\LaTeX/acl2017.sty}
\url{/afs/umbc.edu/users/f/e/ferraro/pub/sty/acl17-word/acl17-word.docx}

Be sure to cite appropriately and follow all academic honesty standards. You may include figures (your own, reproductions, or copies of existing figures); be sure to provide appropriate credit for the figures. However, make the figures count: do not include them simply to pad the paper. Do not consider just “recent” papers; try to find papers from the past 25 years.

Where to Start You may analyze any papers read in class or as part of the assignments.

Google Scholar is an easy way to find linked and cited papers. Another great resource is the ACL Anthology (\url{http://aclanthology.info/}) archives papers by conferences (e.g., ACL, EACL, NAACL, NAACL), journals (CL, TACL), and workshops by year\footnote{Paper ids generally have the form $XY-YZZZ$, where $X$ is a single letter identifier (P is the main ACL, D is for EMNLP, Q is for TACL, etc.), $YY$ are the final two digits of the year (2017 → 17), and $ZZZZ$ is a per-proceedings identifier.} It also offers multiple custom searches: for example, searching “distributed representations” returns papers for crosslingual word representations (C12-1089), representations for relational patterns (P16-1215), and representations for semantic role labeling (D15-1295).

The AAAI digital library also offers an extensive listing of AI-based conferences and proceedings. Of particular relevance are the flagship AAAI, ICML (International Conference on Machine Learning), and KDD (Knowledge Discovery and Data Mining) proceedings. Papers from NIPS (Neural Information Processing Systems) often tend to the more theoretical, but with a decided focus on neural networks.

This paper provides you immense leeway to relate current NLP methods to areas or topics you are interested in. Workshops often offer targeted application and interest areas. There are also special interest groups, which you can find on the main ACL Anthology page. For instance, are you interested in NLP for the humanities, Semitic languages, or biomedical applications?

You are welcome and encouraged to come talk with me, either during office hours, over email, or by appointment to discuss topics, advice on finding relevant papers, and the direction of your paper.

The following is a very small listing of potential starting papers:

7. Rosenfeld (2004): “Two Decades of Statistical Language Modeling: Where Do We Go From Here?”