I'd like as many citations as you can find for me to support the assertion that inter-annotator agreement is probably an upper bound on possible system performance in natural language processing. Please give me the citation and a text snippet that demonstrates that the citation does make this claim.

Natural language processing (sometimes also known as text mining or computational linguistics) is the use of computers to process language in some way, such as finding names of businesses, rating reviews as positive or negative, summarizing news stories, etc.

Natural language processing is often done with an approach called machine learning. Machine learning is a set of techniques for letting computers "learn" for themselves how to classify things (e.g., is this word a name or not? Is this review positive, or not? Does this sentence belong in a summary of this news story, or not?), versus having humans write explicit rules for the computer about how to classify things.

In natural language processing in general and machine learning in particular, we often use data that has been labelled by humans with the correct answers. These humans are called "annotators." For various reasons, we often compute the agreement between the annotators--if two annotators look at the same things, how often do they agree about the classification?

I was told in grad school that it is probably not possible for a computer program that does natural language processing (and probably machine learning more broadly) to perform at a rate that is higher than the inter-annotator agreement for the task. Do I have any good citations for this assertion? No. Please find me as many as you can. I'd like the citations, along with text snippets from the citations that show that they're really making this assertion. Bonus points for Wonder if you give me the citations in BibText format!