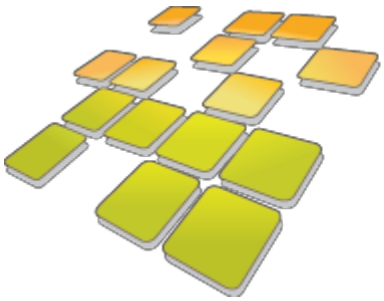

NLP TECHNOLOGIES

Natural Language Processing and the News Media

July 30 2015

Presented to:

The Norman Lear Center

USC Annenberg
School for Communication and Journalism

Contact: Atefeh Farzindar PhD, CEO, NLP Technologies

Table of Contents

A. Introduction

Natural language processing (or NLP, as it is commonly abbreviated) refers to the use of computers to analyze, process and produce natural language for any number of purposes. NLP is a young, interdisciplinary field that draws on computer science, linguistics and artificial intelligence, and its goal is to endow computers with many of the same linguistic abilities that we as humans enjoy. To take one simple example, NLP seeks to allow us to interact with the computers that are increasingly pervasive in our lives using a natural human language like English or Spanish, rather than an artificial computer language like Java or C++. As such, NLP has an important role to play in human-computer interaction.

Language is such a deep-rooted facet of our human nature that we do not generally realize how varied and complex our every-day linguistic abilities are. Consider this last sentence, which you've just read and interpreted with little or no effort. What exactly went into this accomplishment? Abridging somewhat, you first had to distinguish certain of the marks on your screen and recognize them as letters of a particular alphabet. Then you grouped those letters into words belonging to the lexicon of a given language, abstracting perhaps from certain inflected forms; after which you attributed grammatical functions to those words and somehow assigned them and the full sentence a coherent meaning.

Now suppose we wanted to have a computer replicate this feat. What resources could we marshal today that might enable it to do so? In order to decipher the marks on a screen or a page, we would likely call on an **optical character recognition** (OCR) program, which would probably include a **tokenization** module to segment those characters into words and sentences. To determine which language we were dealing with, we could invoke an **automatic language identification** program, and then perhaps a **lemmatizer** to be able to look up the words of that language in a **machine-readable dictionary**. Determining the meaning of the sentences is a more daunting task, but in all likelihood we would begin by invoking some kind of **parser** to group the words into phrases and assign them a grammatical function.

All the programs highlighted in the preceding paragraph are examples of language technologies that form part of the general field of NLP. Each is a focused application that seeks to

automate a particular linguistic ability, and some (particularly those early in the processing chain) can today achieve impressive levels of performance. Needless to say, there are many other language technologies corresponding to other of our myriad linguistic abilities, some of which will be examined below. But first, let us attempt to group the varied programs and components that make up NLP into a few broad and intuitive sectors.

B. Classifying language technologies[1]

1. Text proofing

Though not very common, the term 'text proofing' is used here to designate what used to be called proofreading, i.e. the verification of a text by a human proofreader in order to detect and correct spelling or grammar errors before the text's publication. Text proofing has the identical goal, except that it is achieved by semi-automatic means; which is to say, using a computer program called a spelling and grammar checker. However, while the automatic text proofing programs that are now standard components in word processors and other text editors can detect most spelling errors in a natural language text, and many grammatical errors as well, they cannot as yet reliably *correct* the errors they flag without some human intervention.

Spell checkers were among the first plug-ins to word processing programs, seeing that conceptually they were very simple in design. They essentially consisted of an extensive word list, which in principle included the entire lexicon of the language, and a look-up program that checked every word form in a text against that list. Words in the text that weren't found in that list were flagged as potential spelling errors. For languages like English which have limited morphology, this works relatively well. For more highly inflected languages like German or Turkish, the spell checker needs to include a more complex morphological analyzer; otherwise, the word lists can run up into the millions of word forms.

There are two reason why this simple technology cannot fully emulate the human

[1] Note that there is nothing definitive about the following classification, which is inspired by the taxonomy presented on the website of the NLP group at Sheffield University. It simply aims to group together various branches of NLP which share certain common traits or objectives.

proofreader and correct, as well as detect, common spelling errors. The first has to do with the fact that human languages are not static objects but are constantly evolving; and among the ways they evolve is by coining new words (technically called neologisms) and by importing words from other languages. In both cases, such forms won't appear in the spell checker's word list and so will mistakenly be flagged as a potential error. The second, more common source of spell checker errors are homophones, i.e. two words that sound alike but are spelled differently, and so are frequently confused: *their* and *there*, *see* and *sea*, *hear* and *here*, and a host of others. And here, because both forms appear in the checker's word list, the program won't flag an error when it should. The former problem is sometimes termed noise, the latter problem silence. More recent spell checking programs attempt to tackle the latter by going beyond a simple word list and verifying each word within a limited context of neighboring words.

Grammar checking presents a far more difficult technical challenge than spell checking, requiring as it does a full grammatical analysis, or parse, of each sentence in a text. What humans may not realize, but which comes to the fore when automatic parsing is attempted, is that ordinary, well-formed sentences can often yield hundreds, and sometimes thousands of ambiguous parses – a problem that is only amplified when tackling *ill-formed* sentences. This is why the grammar checkers in most commercial products limit themselves to a simple subset of grammatical errors which can be detected on a strictly local basis, e.g. number, person and gender agreement between verb and subject, or between noun and adjective in Romance languages. It is also why users frequently switch off grammar checking in their applications.

Another somewhat novel application of text proofing is the automatic detection of text reuse, one obvious case being plagiarism, or the unauthorized reuse of a text. >> In collaboration with the Department of Journalism at the University of Sheffield, the NLP Group at that institution created the METER corpus, expressly for the purpose of studying of text reuse. METER consists of a large collection of news stories written by the Press Association, a major UK press agency, along with related stories that appeared in nine major British newspapers. Using the labels provided by their colleagues in the Journalism Department, researchers in the NLP Group trained a classifier that was able to distinguish with a high degree of success stories that were wholly derived from the PA newswire versus stories that were entirely original. (Stories that

were partially derived from the newswire proved more difficult to identify.) Subsequent work by researchers at the University of Darmstadt improved upon those results by factoring text **structure** and **style** (in addition to content) into the text similarity measure.[2] 2

2. Speech processing

Given our focus on media and, in particular, on the texts of news stories drafted by the media, we won't go into great detail on this important branch of NLP, restricting ourselves instead to some brief remarks on its two major sub-domains.

Speech is the most natural medium of human linguistic exchange – many of the world's languages still don't have writing systems – and so it is only natural that we should want to speak to our computers and be understood by them. The enabling technology here is called **automatic speech recognition** (ASR), an area in which impressive progress has been achieved in recent years. These days, none of us think twice about picking up the phone and having a conversation with an automated system, which asks what we want and usually interprets our spoken responses correctly. We can as well issue vocal commands to our computer's operating system, or to the cars we drive, or our mobile phones. Automatic dictation systems (a.k.a. **speech-to-text**) are also gaining in popularity, even among finicky translators. Until recently, the basic trade-off in ASR used to be between systems that were open to a large number of speakers but only handled a small vocabulary, and large-vocabulary systems that required extensive training and tuning to each user's pronunciation. This is less and less the case today. The Google voice typing system that is now standard on Android phones requires no training, and is impressively accurate right out of the box. Furthermore, it is available in about 25 languages, some of which (like Spanish and Arabic) also distinguish numerous regional dialects.[3]

---

[2] For more on this, see http://nlp.shef.ac.uk/research/reuse-anomaly.html

[3] All of which is testimony to the remarkable success of the statistical corpus-based techniques that revolutionized ASR in the 1980's, and which served as a model for a comparable revolution in statistical machine translation.

As for **text-to-speech**, its applications are becoming more and more widespread, one of the best-known being the synthesized voice that gives us directions on our GPS devices.

3. Information access

Unlike the two previous sectors considered above, information access refers to a much larger and more heterogeneous domain of NLP. What its various branches have in common is this: they all process large amounts of unstructured textual data in order to facilitate access to various forms of information.

Information retrieval is undoubtedly the most widely used and best known exemplar, even though most people probably don't realize that this is what they're doing when they routinely submit their queries to search engines like Google or Bing. Just how is NLP involved in what is surely among the most common of all activities on the Web? Without our knowing it, these search engines automatically subject the queries we unreflecting dash off to a number of important linguistic processes. Suppose, for example, that I want to know which is the most popular search engine on the Web, and so I submit the following query: "most popular search engine". Google will take my query and expand it, i.e. add to it an inflected form that I did not explicitly specify; and consequently, in the search results, I will find links to pages that include information on popular search engines (in the plural), as well as links to pages that talk about the one most popular engine. And similarly if my query contains the inflected form of a verb; the engine will expand the query to look up pages containing that verb's other forms. In both cases, basic linguistic processing is involved; to wit, morphological analysis. Or suppose I inadvertently misspell one of the terms of my query, submitting, for example: "most popular search engines". Google will automatically correct that error and inform me on the results page that it has queried "most popular search engine" instead (albeit offering me the option of resubmitting my original query). At first glance, this may appear to involve no more than the application of a spell checker to the user's queries. The resource that Google is consulting, however, is more than just a list of the language's lexicon. For one thing, the search engine will not balk at or attempt to correct well-formed proper names that appear in many queries. And for another, Google now auto-completes our queries; i.e. consulting an enormous list of the most frequent requests it handles

every day, it suggests the full form of several possible queries before we've finished typing ours.

Now notice the form of the information that such search engines return. For the most part, the results page is populated by a list of links to Web pages that will hopefully contain the information that we are looking for. But to confirm that hypothesis and actually obtain the information that interests us, we need to click on the links, access the web page in question and then read the text on that page. Let's backtrack now to that 'for the most part' proviso. Along with each search result, Google also furnishes a short piece of text – it currently has a maximum length of 156 characters – which is intended to provide a short summary of the information contained on the associated Web page. And indeed, reading through these short snippets (as they are called) is often sufficient to allow us to decide whether it is worth our while to visit the Web page in question. How these snippets are actually generated need not concern us here. The important point is that **automatic summarization**, in the form of snippets or as applied to many other forms of text, provides yet another example of an NLP technology that we have come to rely on, almost unconsciously. Seeing that information overload is becoming an increasingly acute problem in so many walks of modern life, the ability to automatically generate accurate and reliable summaries clearly corresponds to a pressing need. Producing such summaries, however, is an extremely challenging problem which would seemingly require an in-depth understanding of the texts that need to be condensed, although sub-optimal solutions are of course possible. The NLP community has organized specialized conferences dedicated to encouraging R&D in this area and has developed particular evaluation techniques (notably, the metric known as ROUGE) to assist in the objective comparison of various algorithms which can be applied to different summarization tasks, e.g. single vs. multi-document summarization; extractive vs. abstractive summarization.

At the beginning of this section, we employed the term '**unstructured textual data'** without explicitly defining what we meant by it. The key word here is *unstructured*: unstructured textual data refers to all types of text that do not conform to any pre-defined data model, or are not organized in any pre-defined formal manner. The text in this report, for example, is unstructured, conforming to no predictable format that might help a computer program process and interpret it; and of course, so is the overwhelming majority of the great masses of text that

appear on the Internet. The branch of NLP that aims to process huge quantities of text – more than any human could possibly digest – and extract from it particular types of information, which may vary according to different users' needs, is called, naturally enough, **information extraction**, or IE. And one well-known variety of IE is **named entity recognition**, or NER. These are programs that are designed to scan large quantities of unstructured text and automatically identify different types of named entities that appear in the text, such as the names of persons, places, organizations, as well as times and dates, etc.; and once they've identified these named entities, the programs extract and copy them into a pre-defined formal data structure. IE has obvious applications in the world of military security and intelligence; but more recently, it also pops up in such inoffensive applications as Apple Mail, where it is used to identify dates, times and locations, and automatically help populate the user's calendar. Existing NER programs are employed to extract names and places and to identify stories that mention a given set of 'influencers'. In fact, intelligence agencies combine such programs with speech recognition to do perform these tasks with spoken news broadcasts. As in the case of automatic summarization, there are specialized conferences and evaluation metrics devoted to IE.

In our brief discussion of IR above, we mentioned two forms of information, or types of responses to queries that search engines currently return: a list of URL's and a short summary of the information on each of the listed Web pages. More recently, a third type of response has begun to appear, at least for certain factual-type questions that users may submit to Google or to Bing. If, for example, I submit the following question to these engines – "Who was the father of Queen Elizabeth II?" – both Google and Bing will display the correct answer – George VI – in a box at the top of the results page. Moreover, if I dictate the same question to Google on my Android phone, Google will speak the correct answer: "Her father is George VI." Not surprisingly, this type of IR is called **question-answering**, or QA, and it was dramatically brought to the public's attention in 2011, when Watson, IBM's question-answering system, defeated two former champions on the TV quiz show *Jeopardy*.

The unreflecting ease with which we humans respond to such questions tends to mask the enormous complexity involved in getting computers to approximate the same behavior. We cannot go into the technical details here – books and scores of articles have been written on the

subject – but an indication of the difficulty of the *Jeopardy* challenge is the fact that dozens of top researchers at IBM and a score of universities worldwide devoted themselves to the task for over five years. One source of the difficulty comes from the fact that Watson is an *open-domain* QA system. Questions on *Jeopardy* can and do refer to any topic imaginable. ***Closed-domain*** systems in principle have an easier task, and can often exploit domain-specific knowledge that is structured in formalized ontologies. Watson, on the other hand, had access to huge quantities of both structured and unstructured information, including dictionaries and thesauri to assist in the linguistic analysis of the questions, enormous gazetteers, encyclopedias like Wikipedia and DBpedia – in all, more than 200 million pages of information, stored on four terabytes of disk space. And perhaps most remarkably, the system was able to produce its answers in real time, no more slowly than the human contestants.

Watson was an exploit, designed to illustrate the potential of QA technology and stimulate the development of more practical applications. And here too, the project has been a success. Less than two years after the *Jeopardy* show, IBM announced an agreement with the Sloan-Kettering Cancer Centre to develop an application of the Watson technology to the medical domain; and agreements with other medical centers have followed. Progress in automatic question answering has indeed been impressive in recent years, although much remains to be done, particularly in the area of **open-domain** QA. This can been seen by returning to the query that we submitted to Google and to Bing at the beginning of this section. If instead of querying "Who was the father of Queen Elizabeth II?", I submit the question "Who succeeded George VI on the throne of England?", Bing provides no short answer but only a list of URL links, most of which refer to articles on George VI. Google fares only slightly better, providing as its short answer a paragraph drawn from an article on the British royals that mentions the daughter of George VI, who later became Queen Elizabeth II.

Another branch of Information Access that we have yet to consider, which is far less difficult than question-answering, is **document classification**. The challenge here is not to *extract* information from one or more documents, but rather to classify a stream of incoming documents into one or more classes of a pre-established set of categories. For example, given a press feed or other type of newswire, a document classification system might be used to route the different

stories to different news desks, based on their content. One common way of doing this is to use a technique called a Naïve Bayes classifier, which models document classes as an independent set (or bag) of words. Incoming documents are analyzed and compared to the parameters defining each document class, in order to determine the probability of its belonging to one or another category. Though the two may appear to be similar, document classification is not identical to **document clustering**, where the task is to partition a given set of documents into groups (or clusters) based on certain shared properties. The training involved in classification is supervised; it relies on a hand-labeled corpus of correct examples; the training involved in clustering is non-supervised.

4. Natural Language Understanding (NLU)

If the NLP tasks in the previous section were primarily concerned with extracting or obtaining information from large bodies of unstructured text, the focus of the tasks in this section is on getting computers to understand part or all of the *content* of such texts. One very simple example is provided by **automatic language identification**. Faced with a text in a language not his own, a user – and even more so, a computer program – needs to know what language the text is written in before he (or the computer system) can do anything with it, such as forward it to the appropriate translator or machine translation system. Automatic language identification programs are designed to provide just that information, i.e. they can determine what language a text is written in, using methods that are often similar to those employed for text classification, only at the character level. And as long as the text to be identified is of a certain minimum length – sometimes no more than 50 characters – these programs tend to be very accurate.

Let us now consider a significantly more challenging task for NLU. Suppose that, as a baseball fan, I have heard that the best bats come from Kentucky, and I want to confirm that contention. I therefore submit the following query to Google: “bats from Kentucky.” Lo and behold, the entire first page of results returned by that search engine – in fact, the first 15 links – have nothing to do with baseball, but are all about the nocturnal flying mammals known as… bats. And that, of course, is because the word form ‘bat’ is ambiguous in English, i.e. it has (at least) two very distinct senses: one referring to the rounded piece of wood swung by hitters in baseball, and the

other referring to the nocturnal flying mammal which happens to be found in large numbers in the caves of Kentucky. In NLP, the task of resolving this type of ambiguity is known as **word sense disambiguation**, or WSD, and our very simple example illustrates why it is so important for IR: without it, search engines will often overwhelm users with irrelevant results. One of the reasons why WSD is such a difficult problem is that word senses, the entities with which it has to deal, are rather nebulous objects. Dictionaries, whose job it is to define word meanings, very often disagree on the number and the precise definitions they assign to even the most common words of a language. Needless to say, this makes comparing and evaluating different techniques for WSD extremely difficult. It was largely to overcome this difficulty that standardized, public WSD evaluation campaigns like Senseval were first organized in the late 1990s. These competitions feature various disambiguation tasks, e.g. monolingual vs. multilingual disambiguation, approaches using supervised vs. non-supervised methods,[4] etc.[4] And crucially, they provide all participants with a uniform test set, and evaluate each system's performance against the same gold standard, which is usually produced by a number of human judges. The time and cost of producing such materials had been a significant obstacle to progress in this and in other fields of NLP, and such public competitions have done much to advance the state of the art.

So where do things now stand in WSD? The problem remains an extremely challenging one, for which a number of complementary approaches have been proposed in recent years. In 2012, Google announced its own Semantic Search project, which is intended to reinforce its classic keyword search algorithm, in part by integrating WSD techniques. Returning to our "bats from Kentucky" example, what makes this short query so difficult is that it doesn't provide enough context to allow for the disambiguation of the word 'bats'. It's not clear exactly how Google operates, but it's sufficient to add a single modifier to the query, e.g. "*hickory* bats" or "*furry* bats", for the search engine to get things right, all the first page results displayed for the former now pertaining to baseball, while all the results for the latter relate to the nocturnal flying mammals.

[4] Supervised methods use hand-annotated data to train from; non-supervised methods eschew such data.

We stated (a little provocatively) above that word senses are rather nebulous objects. One context in which the different senses of a homographic word like 'bat' may be clearly distinguished is translation. In other words, the two senses of a single word form in one language may often be translated by completely different words in another language. This is the case with 'bat' when it is translated into French: the baseball sense is translated as 'baton' and the flying mammal sense as 'chauve-souris'. Now as it happens, one of the very first applications proposed for the new digital computers which had been developed during World War II was **machine translation** (or MT); and one of the very first arguments against the possibility of fully automatic, high quality machine translation involved lexical disambiguation. In 1960, Y. Bar-Hillel published an article containing a very simple example sentence – "The box is in the pen" – which, he claimed, demonstrated the infeasibility of MT's ultimate goal. The argument hinges on the ambiguity of the word 'pen', which may designate a writing instrument or an enclosure for little children. If the sentence poses no difficulties for humans, Bar-Hillel argued, it is because we know that boxes cannot normally fit into writing instruments. However, for a machine to correctly disambiguate this and scores of similar sentences, it too would require extensive world knowledge and the ability to reason over it – something Bar-Hillel considered altogether unimaginable.

Unfortunately, Bar-Hillel's argument went largely unheeded in the early years of machine translation. MT researchers continued to develop increasingly sophisticated systems, many of which attempted to formalize and render explicit the underlying meaning of the sentences they sought to translate. For as any translator will tell you, you cannot translate a sentence without first understanding what it means. While this is certainly true for humans, it proved to be something of a dead end for MT. As the worldwide demand for translation continued to grow exponentially, the analysis-based MT systems that were developed up until the late 1990s were simply not up to the task, and very few found large-scale commercial applications. At the end of that decade, however, MT began to undergo a radical change, with the emergence of a new datadriven, statistical paradigm. Instead of relying on linguists and lexicographers to hand-code complex transfer rules, system developers began to apply powerful machine-learning algorithms to large corpora of previously translated texts, thereby enabling the machines to automatically

infer translation correspondences directly from the data. And it didn't take long before these new statistical MT (SMT) systems began to outperform the traditional rule-based systems in open competitions sponsored chiefly by the US government. Today, virtually all new MT systems adhere to the SMT paradigm, among other reasons, because such systems are so much less costly to develop. Google Translate is without a doubt the best known of these. In 2013, it was said to handle the astonishing figure of a billion translations per day.

While it is indisputably true that machine translation has made remarkable progress since the emergence of the new statistical paradigm, it must be said that we are still quite far from the field's ultimate goal, which Bar-Hillel formulated as fully-automatic, high-quality translation.[5] That said, what is interesting for our purposes is that these new state-of-the-art systems, unlike their rule-based predecessors, do not appear to rely on any explicit representation of the meaning of the texts they aim to translate, i.e. NLU no longer seems to play an important role in MT. In fact, many of the automatically inferred translation correspondences that are stored in an SMT's translation table will make no sense to a human linguist and will in all likelihood be ignored because of their low probability scores. In this regard, SMT seems to represent a serious departure from artificial intelligence's traditional approach of programming machines in such a way that they emulate the behavior of human experts. SMT systems now translate between a wide array of language pairs, they often translate impressively well, but they do so in ways that bear little or no resemblance to the way human translators operate (insofar as we understand the latter).

Before leaving the area of NLU, let us briefly consider one final sub-domain which has found interesting commercial applications in recent years. **Sentiment analysis** is concerned with understanding one particular aspect of text content, to wit, the affective attitudes (generally, positive or negative) that are expressed in a piece of written text, and in aggregating those results with many others in a form that users can easily assimilate. Early sentiment analysis programs relied heavily on hand-crafted polarity lexicons, which simply listed positive and negative terms

[5] Indeed, the great majority of the daily translations produced by Google are not intended for publication but are simply used for gisting, or information gathering purposes.

and then tallied up their occurrences in a text to produce a cumulative evaluation. One of the first applications was to movie and book reviews, and while the technology was somewhat simplistic, these applications proved quite popular. The explosive growth of social media in recent years, and with the ensuing availability of 'big data', has stimulated the development of more sophisticated approaches to what is also called opinion mining. Many businesses, from large multinational corporations to local restaurants, are keenly interested in learning how clients react to their products and services, and there is now no shortage of tech companies offering this kind of Web monitoring services. A good illustration of the current state of the art is provided by the products search services that engines like Google and Bing now offer. Not only do these compare prices of different vendors for a given product; they also automatically identify the pertinent attributes for each product and aggregate the consumer evaluations expressed in their comments into a five-point rating. To judge by the explosive growth of this sector in the last years, one might be tempted to think that they have reached a state of advanced reliability. Unfortunately, this does not appear to be the case. Although much depends on variables like text type and text length, several recent papers on the evaluation of commercial sentiment analysis programs point to rather disappointing results, i.e. accuracy in the 60% range.[6] Extracting affective attitudes from random unstructured text is clearly a more difficult challenge than tabulating the number of clicks on a Like button.

C. NLP and the news media

Having surveyed the broad sectors of natural language processing, let us now turn our attention to the media, and in particular to the news media, since the aim of this study is to explore the application of various NLP technologies to media newsrooms – those of traditional printed newspapers, those of broadcast media (radio and television), as well as those of new digital media.

The central activity of the news media is to produce news stories, i.e. texts in the form of

[6] See, for example, Cieliebak, Dürr and Uzdilli, "Potential and Limitations of Commercial Sentiment Detection Tools", Zurich University of Applied Sciences, 2013.

articles or dispatches written by journalists in which they present and/or analyze topical events or social and political issues. Regardless of whether they are ultimately printed in newspapers or displayed on a Web site, or spoken on a television or radio broadcast, the important point is that all such news stories are **texts,** and they are generally produced under stringent time pressures. (Few people are interested in old news.) As such, it is only natural to look to various NLP techniques, since these offer the possibility of automating, in whole or in part, the production of such specialized texts, thereby relieving media employees of otherwise tedious tasks or helping them produce their news stories more rapidly and efficiently.

In the first section below, we will provide a few examples of tasks involving the news media where NLP has already been employed with some degree of success. Then, in the section that follows, we will consider other newsroom operations where more advanced NLP techniques could potentially provide much-needed assistance.

1. Well-established applications of NLP

Not so very long ago, all major newspapers used to employ full-time, in-house proofreaders whose job it was to read copy before it went to press, detect any errors of spelling, grammar or style, and correct them. This is much less so today.[7] Not that the news media have entirely abandoned the task of proofreading. Rather, a good part of the work has been delegated to computerized **text proofing** programs, i.e. spelling, grammar and style checkers which generally work in tandem with a human proofreader or copy editor. On their own, the automated text proofing programs may not necessarily be more accurate than a human proofreader. The advantage they offer is speed: they allow for the verification of far more text in much less time than any human could possibly do.

Let us now consider a far more challenging task for NLP in the news media. **Closed captions** are text segments that are inserted into a video broadcast, primarily to allow people with hearing impairments to follow the audio portion of the broadcast which they would otherwise miss. By

[7] Proofreading is still standard practice in scientific journals, but even here it tends to be done by freelancers. Translation agencies that pride themselves on high quality still tend to employ full-time in-house proofreaders.

some accounts, around ten percent of the population in North America suffer from some form of hearing impairment, which is why governments in Canada and the United States passed legislation in the 1990s obliging TV networks to include closed captions in their programs. For television shows that are not broadcast live, closed captioning is usually done by stenographers, in much the same way that court transcripts are recorded, except the stenos may be linked to specialized equipment which automatically converts their abridged code into standard orthography and inserts it into the video file. The problem, however, is there are just not enough competent stenographers to cover the large number of TV shows that require captioning; and training a stenographer is said to take several years. Moreover, live programming, like most news broadcasts, is less amenable to this approach, in large part because it frequently contains new terms and proper names that stenographers and their conversion software have trouble handling correctly. This has led to the development of an innovative way of producing closed captions that relies on automatic speech recognition. But instead of having the ASR system try to decode the anchor man's speech directly, re-speakers (sometimes called 'shadow speakers') are employed to repeat the journalists' narrative in a controlled noise-free environment, thereby achieving impressively low error rates in time delays that still remain altogether acceptable (no more than 4 seconds). Moreover, some of these real-time captioning systems include an off-line Web mining component that regularly scans written new reports so as to identify and add new proper names to the ASR vocabulary. In Canada, this technology is now being employed by one major French-language network for all its news broadcasts, as well as by the Canadian Parliament for its transmission of House of Commons debates.

Closed captions are sometimes confused with subtitles, since both insert text segments onto a video medium in order to convey the spoken dialogue. There is an important difference between the two, however: closed captions are in the same language as the original audio track, whereas subtitles are in another language, and so involve translation. The need to translate newscasts as well as printed news stories has, of course, long been recognized; indeed, it was one of the early justifications for investing in **machine translation**, and even today it remains of vital concern to the intelligence and security establishments. Nowadays, we think nothing of

looking for a particular news item on the Web by submitting a query to our favorite search engine, and then clicking on the 'translate this page' button next to each result in order to obtain a Google or Bing translation in our mother tongue. Yet things were not always so easy; when large daily newspapers first began to publish a Web version, these free online translation services were not available. Recognizing the business potential of being able to offer their news content to non-English speakers, the Los Angeles Times, the second largest daily in the US, launched a project with Alis Technologies in 1997, whereby French and Spanish readers of the paper's online version could request an instant translation of certain stories. Unfortunately, the project didn't last very long, for reasons that were never made public. (This was before the widespread adoption of SMT, and perhaps the quality of the raw machine translations simply wasn't good enough.) Be that as it may, the ability of the news media to attract more readers by offering them stories in their own language, remains as compelling today as it was back then.[8]

To conclude this section on existing applications of NLP to the news media, let us consider a problem that definitely involves the media but is not limited to them. The problem is **information overload** – exposure to more information than we can effectively process – and it may or may not be a new phenomenon.[9] What is certain, however, is that in today's information age, as a result of globalization, the expansion of the Internet, the pervasiveness of email and the explosive growth of the new social media, the quantity of information now available to us is simply staggering, to the point that it may actually hamper our ability to make correct, well-informed decisions. Decision-makers in government, corporations and academia desperately need some reliable way of zeroing in on the information that is of particular importance to them, without wasting their time floundering about in a sea of irrelevant and unreliable data. As we saw above in our discussion of document classification, well-known NLP techniques can help provide part of the solution to this dilemma. Companies have sprung up that offer to monitor the constant flow

[8] Why else would the Huffington Post publish in six different languages (as of 2014), with plans to launch another three linguistic versions?

[9] The term itself was coined by the American futurologist Alvin Toffler in 1970.

of information from news agencies and other sources like popular Websites and Facebook pages and filter them according to the particular interests of their subscribers. Eureka.cc is one such service, marketed by the Canadian company CEDROM-Sni. Increasingly, media monitoring and media analysis, which is akin to sentiment analysis, are bundled together; this is what the company Media Miser does, for example. Both CEDROM-SNi and MediaMiser provide media monitoring services; that is, they scan a broad range of media outlets (newspapers, TV & radio, the Web, the social media, etc.), typically for business clients and large organizations, searching for mention of their client's name, after which they collate and analyze the gathered information in a form which is sometimes referred to as actionable business intelligence. Content curation services like Scoop.it offer another, more collaborative approach of coming to terms with the same information overload problem. Scoop.it, offers a new form of social publishing platform that combines human-mediated content curation with what the company calls "big data semantic technology" to allow individuals and companies to achieve greater impact and visibility for their blogs and websites.

2. New applications to internet journalism

In the days when print dominated the news media, circulation – i.e., how many people subscribed or bought the newspaper – was the dominant factor that determined advertising revenue. And similarly for television and radio, where listener ratings played the same role. In the new age of the Internet, things are not so very different; except that the number of competing news sources is much greater, and the manner in which their readership is calculated has changed. Some readers have their preferred news sources, to which they return every day. But very often, Internet users first turn to a search engine when they want to find news stories on a particular topic. And here, the key to attracting readers is to have the link to one's Web site among the top results on the search page; for it is a well-established fact that the higher the rank on the search results page, the more visitors that page is likely to receive. (Of course, this is not particular to news sites, but applies as well to all types of searches.) In the last fifteen years or so, a new industry has emerged called **search engine optimization**, or SEO, whose goal it is to guarantee just that result, i.e. to ensure that its client's Web site appears at or near the top of the search page results. Just how SEO practitioners obtain this competitive advantage for their

clients is something of a trade secret, and it is frequently claimed that not all of their techniques are legal or ethical. Be that as it may, we understand enough of how the major search engines rank their results to be able to infer some of the SEO techniques that have been employed. The content of a Web page may be edited so that frequently searched (or trending) phrases appear there, or in the page's meta-data; or more dubiously, certain keywords or their synonyms can be hidden in a page's graphics and repeated numerous times, once again with a view to increasing a page's ranking. In fact, ever since the inception of SEO, there has been an ongoing game of cat and mouse between its practitioners and the developers of the major search engines, with the latter attempting to neutralize many SEO techniques in the interest of 'maintaining a level playing field'. Without taking sides in this debate, we simply want to point out that several key SEO strategies crucially call upon information that NLP can help provide, e.g. establishing synonyms and paraphrases, and determining the salience or frequency of one phrase over another.

One publication that has certainly mastered the art of SEO in the new age of Internet journalism is The Huffington Post. According to comScore, the site receives more than 200 million unique visitors each month, making it the most-visited digital-native news site in America. On June 30, 2015, the New York Times magazine published an in-depth story on the Post and its founder, Arianna Huffington. That story, and another on the new digital journalism that recently appeared in the New York Review of Books, focus on some of the factors responsible for the Post's success. Among the factors mentioned in the Times article is the "radical data-driven methodology" which the Post has brought to its home page, "automatically moving popular stories to more prominent spaces and A-B testing its headlines."[10] Another of its core innovations is the large-scale use of aggregation, i.e. the recycling of content from other sources, including the Associated Press, Reuters, the New York Times and the Washington Post. In terms that are not terribly flattering, the Times article describes Ms. Huffington's penchant for hiring young journalism graduates from Yale and paying them a pittance to churn out summaries of stories in

[10] In A-B testing, two versions of a Web page are shown to the members of a focus group in order to determine which of the two they prefer.

other publications.[11] Here is an area where NLP technology will no doubt be playing an increasingly important role in the coming years: **automatic summarization** – not to replace those young Yale graduates, but hopefully to make them more efficient and productive.

Another area of NLP that is likely to become increasingly important in coming years, not just for journalists but for the public in general, is **cross-language information retrieval**, or CLIR. As Marshall McLuhan forecast half a century ago, the world has indeed become a global village, and very often the information that we require is not available in our own native language. Suppose, for example, that I am a journalist researching a human interest story on kinship terms in other languages and cultures. If I submit the following query to a search engine – "kinship terms in different languages" – all the results I obtain will be linked to Web pages in English. Suppose, however, that I have heard of some expert in the area who happens to write in French (or Urdu, or what have you). I therefore go into the Advanced Options of my search engine, specify French as the language of the pages that interest me, and resubmit the same query. The results remain essentially unchanged, because the search engine seeks to match the words of my query with the words of the Web pages. This is where CLIR comes in: it translates either the words of the query, or the words of the Web pages, so that the two intersect, thereby allowing for the retrieval of pages in 'foreign' languages.[12]

D. Conclusion

We began by defining the broad sectors and principal subdomains of natural language processing; we then described a number of well-established applications where NLP is already being successfully employed in the daily operations of media newsrooms; and finally we pointed to several other applications of more advanced forms of NLP technology (some of which are still in the research stage) where NLP could potentially make a significant contribution to newsroom operations, particularly those of the new digital media.

---

[11] Segal, David. (2015, June 30) "Arianna Huffington's Improbable, Insatiable Content Machine", *The New York Times.* Retrieved from: http://www.nytimes.com/

[12] Google used to offer an advanced search feature that performed CLIR. It translated the user's query and retrieved pages in foreign languages which the user selected. Unfortunately, it was discontinued in 2013.

People who currently work in the news media might be tempted to think that this is all well and good, everybody's in favor of progress, but this talk of NLP really doesn't concern them very much. That would be a mistake. On June 28, the American Society of News Editors released its census figures showing that the number of journalists and editors working in the newsrooms of its members declined by 10.4% between 2013 and 2014 – a loss of 3800 jobs in one year! To be sure, the Society tends to focus on print media and its members do not include every publication in the USA. Still, the data shows a clear 10-year trend which does seem to be accelerating, particularly in medium-sized papers.[13] The decline in print advertising has also been well documented, and has yet to be fully compensated by ad revenues on digital platforms. All of which is to say that the news media are going to have to do more with less in the coming years. NLP technologies like those we have described here may well offer them some means for doing so.

[13] The largest papers in the US – those with a daily circulation of more than 250,000, like the New York Times and USA Today – have not (as yet) cut back their newsroom staff.