BASIS TECHNOLOGY

www.basistech.com
info@basistech.com
617-386-2090

careerbuilder®

Case Study

fast A Microsoft® Subsidiary

Solr

RBL ROSETTE Base Linguistics

CHALLENGE

CareerBuilder, the global leader in human capital solutions, operates the largest job board in the U.S. and has an extensive and growing global presence. The CareerBuilder.com content is the very definition of Big Text: mountains of structured and unstructured text data (resumes and job listings) in many languages. CareerBuilder's mission is to empower employment, striving to organize the world's human capital data and make it meaningful for society. Fundamental to this mission is delivering highly accurate and reliable search results to match the right people with the right jobs.

The CareerBuilder.com content is the very definition of Big Text: mountains of structured and unstructured text data (resumes and job listings) in many languages.

Previously, CareerBuilder utilized the Norwegian search engine FAST as their core search technology. FAST proved to be an excellent solution for CareerBuilder for years, in no small part due to FAST's reliance on Basis Technology's Rosette® Base Linguistics (RBL) to provide advanced language analysis services such as tokenization, part of speech tagging, decompounding, and lemmatization.

TOKENIZATION
貴社の記者が汽車で帰社した。

DECOMPOUNDING
samstag *morgen*

NOUN PHRASE EXTRACTION
Miami University of Ohio

LEMMATIZATION
am *are* *is* → to be

PART OF SPEECH TAGGING
Pronoun Verb Adjective Noun

SENTENCE DETECTION
" An investment in knowledge pays the best interest."
- Benjamin Franklin

In 2008, Microsoft® purchased FAST and shifted the product's direction toward powering search within their Sharepoint® Enterprise Solution. After this direction shift, CareerBuilder, like many former FAST customers, turned to the open source search engine, Apache Solr™. With this major technology migration, CareerBuilder sought to optimize two different search scenarios to deliver a better user experience: User-initiated search and automated recommendations.

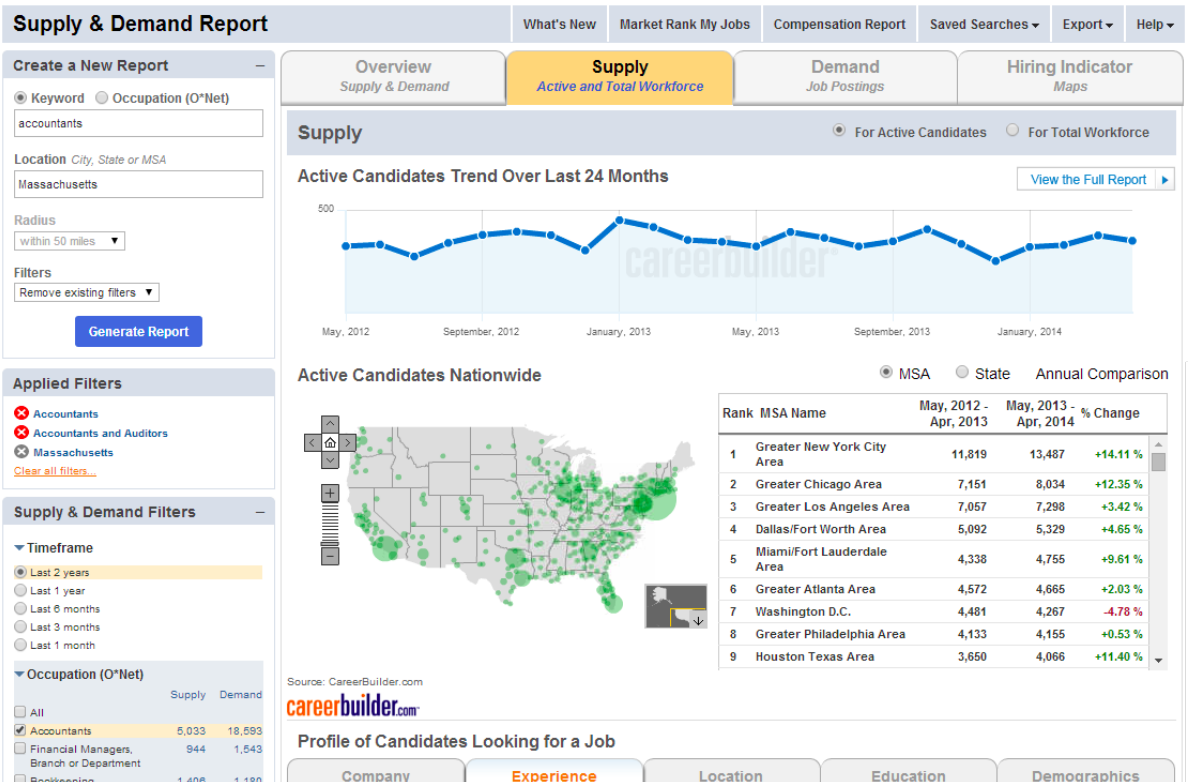
When a user enters a search term, CareerBuilder has to ensure that it is delivering the "right results", i.e. the largest possible set of relevant results (recall), with the smallest possible set of irrelevant ones (precision).

While this seems straightforward when matching exact keywords such as "Java" or "CPR", many search terms have linguistic variations such as "engineer" vs "engineers," where you would want to ensure you always match on all variations of the same root words.

At the same time, you want to be sure that the variations make sense - you generally wouldn't want the term "accountant" to match the term "account", as an account manager would not be a good match for the keyword "accountant". CareerBuilder also provides several search-powered Big Data Analytics products modeling labor market trends (screenshot below), so ensuring that the correct keyword variations are matched is critical to the accuracy of these products.

In addition to traditional keyword searching, CareerBuilder also provides a recommendation engine that is able to automatically find relevant resumes for recruiters (based upon a job or another resume) or relevant jobs for job seekers (based upon their behavior or their resume).

Doing content-based recommendations is dependent upon the ability to extract key terms from one job description or resume that can be used to search for related terms in other documents. A key challenge in doing this well is the ability to distinguish which kinds of terms are important across various languages.



The majority of CareerBuilder's job and resume content is in English, so their highest priority was to deliver the best possible search results for English language queries. However, with CareerBuilder's growing global presence, they needed to provide the same high-quality experience for their users in languages as diverse as Chinese, German, Greek, and Spanish.

SOLUTIONS

As the CareerBuilder search team migrated to the Lucene/Solr platform, they needed to evaluate the ability of Solr to address the language-matching challenges outlined above and whether it was necessary to add advanced language analysis services to improve the out-of-the-box functionality.

CareerBuilder was quite satisfied with the quality of the search results they were getting from FAST, so they used it as their experimental baseline.

The team then set up a head-to-head comparison of several language analysis software solutions to determine their relative impact on the precision and recall of the search results, as compared to the results they received from FAST. Included in the comparison were the base Solr configuration with no additional linguistic support, a standard Solr stemmer called the "Snowball Stemmer", the well-known "K Stemmer" (for English comparison only) and finally, Basis Technology's Rosette Base Linguistics (RBL), which provides advanced lemmatization.

LEMMATIZATION

am *are* *is* → to be

Most search engines utilize a crude method of chopping off characters at the end of a word in the hopes of removing unimportant differences. This method, called stemming, often results in extra recall and poor precision. Instead, RBL finds the true dictionary form of each word, known as a lemma, by using vocabulary, context, and advanced morphological analysis. Indexing the root form increases search relevancy and slims the search index by not indexing all inflected forms. Alternative lemmas are also made available to supplement indexing.

Example: English
Linguistic analysis is useful for every language; lemmatization for English improves recall and precision.

CHALLENGE	QUERY	STEM	LEMMA
Two unrelated words may share a stem.	animals animated	anim	animal animate
Stemming may deliver unintended results.	several	sever	several
Irregular verbs and nouns stump the stemmer.	spoke	spoke	speak (v.) spoke (n.)

The results were definitive. Across the board, RBL delivered results that were the closest to the FAST experimental baseline in terms of precision and overall accuracy. The stemming algorithms generally delivered many more irrelevant results that would have negatively impacted the user experience.

The reason? RBL uses vocabulary, context, and advanced morphological analysis to determine the lemma, or dictionary definition, of the words in the index and in the search query. Therefore the meaning of the words being searched for is more accurately matched to the meaning of the words in the index, and unrelated words with similar spellings are not included in the results.

> "Our experiments showed that most open source stemmers resulted in either too many bad results from over-stemming (low precision) or too many missing results (low recall) compared to a lemmatizer. RBL allowed us to optimize the balance between precision and recall to provide a superior experience for our customers' searches."
>
> *Trey Grainger - Director of Engineering, Search & Analytics at CareerBuilder*

However, the head-to-head comparison did reveal some interesting exceptions that pointed to another one of RBL's features: decompounding.

DECOMPOUNDING

samstag / *morgen*

RBL breaks down compound words into sub-components and delivers each individual element to be indexed. This is especially useful for increasing search relevancy in languages such as German and Korean.

Example: German
Samstagmorgen is a compound word formed with Samstag (Saturday) and morgen (morning). Decompounding allows for an appropriate match when searching for "Samstag".

In languages such as German, Dutch and Greek, compound words are quite common. For example, the word Fließbandproduktionsleiter (assembly line production manager) is a compound of assembly line (Fließband), production (Produktion) and manager (Leiter). In cases like this, if a user searched for "Produktion" they wouldn't retrieve the word Fließbandproduktionsleiter even when using a traditional stemming approach.

RBL is able to decompound words like Fließbandproduktionsleiter and provide relevant matches for the individual words Fließband, Produktion, and Leiter in the index.

The results were definitive. Across the board, RBL delivered results that were the closest to the FAST experimental baseline in terms of precision and overall accuracy.

The results in the head-to-head comparison actually indicated a substantial increase in recall in these three languages without a significant decrease in precision. These results suggested that RBL was finding a larger number of relevant documents that were otherwise being missed by other algorithms that could not properly decompound the words being inserted into the index.

ENHANCED RECOMMENDATIONS

CareerBuilder was also able to leverage the Rosette Language Identifier (RLI) and RBL's part of speech tagging to enhance their recommendation engine. After the language of a source document (e.g. job or resume) is identified and the appropriate lemmatizer is chosen, RBL identifies each word's part of speech (e.g. noun, adjective, verb, etc.).

PART OF SPEECH TAGGING

As part of the lemmatization process, statistical modeling is used to determine the correct part of speech, even with ambiguous words. Each token is then tagged for enhanced comprehension and search relevancy.

三载匆匆，现在的我深深懂得：昨天的成绩已成为历史，未来的辉煌要靠今天脚踏实地坚持不懈地努力去实现。
在这斑斓多彩日新月异的时代，只有培养能力、提高素质、挖掘内在的潜能，才能在激烈的社会竞争中立于不败之地。

Ενεργειακά αυτο-μίζα και τον όγκο παραγωγής, εύκολα προσαρμοζόμενο στις μεταβαλλόμενεςπροτεραιότητες. Εξαιρετικές ικανότητες διαχείρισης πελατών και προμηθευτών σχέσεις, σε συνδυασμό με την μεγάληπείρα στηδιαχείριση των επιχειρήσεων σε μικρές και μεσαίου μεγέθους επιχειρήσεις.

EigenstŠndige Bearbeitung der Kunden von sechs Fachzeitschriften einschlie§lichMahnwesenAufbereitung derAbonnementsentwicklung von sechs Fachzeitschriften und gemeinsameAnalyse mit Controlling undProduktmanagementMitarbeit in einem Projektteam zur Migration desgesamten Kundenstamms auf eine neueSoftwareplattform

Experiencia Laboral en el campo de la contadur'a Pœblica auditoria interna y Revisor'a Þscal.
Gerente y accionista mayoritario de Þrma de contadores pœblicosy Socio mayoritario de empresadecomercializaci—nde confecciones.

This additional tagging allows CareerBuilder to extract only specific parts of speech within a document, such as nouns, which are identified as the most valuable types of words when matching between document. CareerBuilder then applies their own statistical modeling to determine the most relevant of the extracted terms within the document, based upon the frequency of their occurrence across other kinds of documents, and uses these terms (along with some other machine learning techniques) to generate highly relevant recommendations for other documents.

RESULTS

CareerBuilder's success, with over 24 million unique visitors a month, speaks to the effectiveness of their search-based user experience. Clearly they are succeeding at matching the right person to the right job at the right time, delivering a great value to both job seekers and companies.

CareerBuilder is a technology company at heart, with a robust staff of software engineers. Every decision that they make to utilize external software has to be supported by a clear analysis that it is more cost effective to buy something than to develop it themselves. CareerBuilder has been an active customer for the past 4 years and has never regretted their decision to buy Rosette. In fact, RBL acts as a key component within their search infrastructure, allowing them to focus on developing more specialized entity extraction and data analysis engines that yield even greater value to their customers.

> "At CareerBuilder, we are building talent management software that unlocks meaning in unstructured human capital data. Our core competencies include search, data classification, matching, and big data analytics, and relying upon Basis Technology for our linguistic analysis (in over a dozen languages) allows us to remain focused on our core competencies and ultimately provide more value to our customers."
>
> *Trey Grainger - Director of Engineering, Search & Analytics at CareerBuilder*

"relying upon Basis Technology for our linguistic analysis (in over a dozen languages) allows us to remain focused on our core competencies and ultimately provide more value to our customers."