April 26, 2013

Enabling High-Quality Search in European Languages

A Comparison of Stemming vs. Lemmatization for French, German, Italian, and Spanish

Half of all word families in both French and Spanish benefit from analysis with lemmatization rather than stemming.

BASIS TECHNOLOGY

We put the World in the World Wide Web®

ABOUT BASIS TECHNOLOGY

Basis Technology provides software solutions for text analytics, information retrieval, digital forensics, and identity resolution in over forty languages. Our Rosette® linguistics platform is a widely used suite of interoperable components that power search, business intelligence, e-discovery, social media monitoring, financial compliance, and other enterprise applications. Our linguistics team is at the forefront of applied natural language processing using a combination of statistical modeling, expert rules, and corpus-derived data. Our forensics team pioneers better, faster, and cheaper techniques to extract forensic evidence, keeping government and law enforcement ahead of exponential growth of data storage volumes.

Software vendors, content providers, financial institutions, and government agencies worldwide rely on Basis Technology's solutions for Unicode compliance, language identification, multilingual search, entity extraction, name indexing, and name translation. Our products and services are used by over 250 major firms, including Cisco, EMC, Exalead/Dassault Systems, Hewlett-Packard, Microsoft, Oracle, and Symantec. Our text analysis products are widely used in the U.S. defense and intelligence industry by such firms as CACI, Lockheed Martin, Northrop Grumman, SAIC, and SRI. We are the top provider of multilingual technology to web and e-commerce search engines, including Amazon.com, Bing, Google, and Yahoo!.

Company headquarters are in Cambridge, Massachusetts, with branch offices in San Francisco, Washington, London, and Tokyo. For more information, visit www.basistech.com.

BASIS TECHNOLOGY

1. INTRODUCTION

Providing high-quality search results is more important than ever before. Search is now the ubiquitous way that users interact with data sets to get their jobs done. As overseas users continue to create and consume data in their native languages, it is crucial to understand how your search engine handles these languages.

In this paper, we focus on how to provide high-quality search results, using German examples that illustrate the challenges present across many Western European languages, and the technology that makes this possible. Specifically, we will look at lemmatization—a natural language processing (NLP) technique to identify the dictionary form of a word—versus stemming—a naïve alternative.

As illustrated by the examples provided, lemmatization improves search quality in a wide range of scenarios, while stemming has limited utility. This has been confirmed by enterprise search users over many years, as well as corroborated by quantitative analysis, which shows that about half of all word families in the set of the 40,000 most commonly used words in both French and Spanish[1] benefit from analysis with lemmatization rather than stemming in the context of search and other keyword-based applications.

2. ENABLING HIGH-QUALITY SEARCH IN EUROPEAN LANGUAGES

Search quality is a combination of a number of different factors, but here we will focus on measurements called *precision* and *recall*. *Precision* examines the results that are returned, and measures how many of the returned results are correct, given the query. *Recall* examines all possible correct results, and measures how many of them were returned. By thinking of search results in terms of precision and recall, we can better understand the challenges of European languages, and how different approaches deal with these challenges.

In linguistic terms, European languages are highly inflected. This means that words are modified based on tense, gender, case, quantity, aspect, and other factors. In English, this is primarily conjugation of verbs and a number of irregular nouns, but some European languages frequently use declension, or the inflection of nouns, adjectives or pronouns. In short, it is common for the forms of words to change based on how they are used.

To address this challenge, words must be "normalized," at index and query time, to ensure that a query for one form of a word matches the other forms of the word. There are two main approaches that a search engine can use to normalize these word variations:

- Stemming: language-specific rules for removing characters from the ends of words.
- Lemmatization: identifying the dictionary form of a word, based on how it was used.

Lemmatization is the superior way to normalize words from Western European languages because it draws on an understanding of the language and uses context when identifying the correct dictionary form of the word. *Stemming*, on the other hand, uses a simple rule based approach to

---

[1] Based on word frequency counts from Google Books Ngrams for resources published in 1950 and later. Note that some unigrams are not real words (e.g., "■" "a.j.") and, hence, were excluded.

chop characters off of words that does not take advantage of context. In many cases, stemming adds non-words (i.e. stems) to your index. This may cause unrelated words to share the same stem, creating precision and recall problems when indexing and searching.

In the examples that follow, the middle column illustrates some of the problems associated with using stemming algorithms and the final column illustrates how morphological analysis handles these variations correctly.

Compound words are an additional problem presented by languages such as Danish, Dutch, German, Hungarian, Norwegian and Swedish. For compound words—such as the German word "Kinderbuch" (children+book)—decompounding when indexing is crucial for accurate search in these languages. We also include some German examples illustrating compound issues as they relate to precision and recall in search.

3. FRENCH

3.1. Collisions

Identically spelled words with different meanings create the same stems, resulting in decreased *precision*. According to our analysis, this occurs in 14% of French word families.

Input	Stem	Lemma
été (summer)	→ été	→ été (summer)
été (was)	→ été	→ être (to be)

Input	Stem	Lemma
bois (I drink)	→ bois	→ boire (to drink)
bois (woods)	→ bois	→ bois (woods)

Input	Stem	Lemma
aimant (magnet)	→ aim	→ aimant (magnet)
aimant (loving)	→ aim	→ aimer (to love)

Input	Stem	Lemma
badine (he/she trifles)	→ badin	→ badiner (to trifle)
badine (switch)	→ badin	→ badine (switch)

L'été sera chaud. (The summer will be hot.)
Il a été blessé. (He's been injured.)

Je bois beaucoup d'eau avec les repas. (I drink a lot of water with meals.)
Je vais marcher dans le bois. (I will take a walk in the woods.)

J'ai un aimant sur le frigo. (I have a magnet on the fridge.)
En aimant, se libère-t-on de la société? (By loving, does one become free from society?)

One ne badine pas avec le travail. (Work is not to be taken lightly.)
Où est la badine? (Where is the switch?)

3.2. Collisions

Differently spelled words create the same stem, resulting in decreased *precision*. According to our analysis, this occurs in 48% of French word families.

Input	Stem	Lemma
aire (area)	→ air	→ aire (area)
air (air)	→ air	→ air (air)

Input	Stem	Lemma
publication (publication)	→ publiqu	→ publication (publication)
publique (public)	→ publiqu	→ public (public)

Input	Stem	Lemma
cuir (leather)	→ cuir	→ cuir (leather)
cuire (bake)	→ cuir	→ cuire (to bake)

Input	Stem	Lemma
ferme (farm)	→ ferm	→ ferme (farm)
fermes (you close)	→ ferm	→ fermer (to close)

Cette pièce couvre une aire de 12 mètres. (This room is 12 meters square in area.)
L'air est pollué dans cette ville. (The air is polluted in this city.)

J'ai soumis ma publication. (I submitted my publication.)
Ceci est une école publique. (This is a public school.)

Ce manteau est en cuir. (This coat is made of leather.)
Veux-tu cuire un gâteau? (Do you want to bake a cake?)

Il a une ferme. (He has a farm.)
Tu fermes la porte. (You close the door.)

3.3. Failure to Normalize

Words not normalized when stemmed result in decreased *recall*. According to our analysis, this occurs in 6% of French word families.

Input	Stem	Lemma
genoux (knees)	→ genoux	→ genou (knee)

Input	Stem	Lemma
pneus (tires)	→ pneus	→ pneu (tire)

Input	Stem	Lemma
suis (I am)	→ suis	→ être (to be)

Input	Stem	Lemma
a (he/she has)	→ a	→ avoir (to have)

J'ai mal aux genoux. (My knees hurt.)

Il faut changer ces pneus. (These tires need to be changed.)

Je suis ici. (I am here.)

Il a une maison. (He has a house.)

3.4. Inconsistent Stems

Different stems for different inflections of the same word result in decreased *recall*. According to our analysis, this occurs in 9% of French word families.

Input	Stem	Lemma
suis (I am)	→ suis	→ être (to be)
est (he/she is)	→ est	→ être (to be)
sommes (we are)	→ sommes	→ être (to be)

Input	Stem	Lemma
vais (I go)	→ vais	→ aller (to go)
va (he/she goes)	→ va	→ aller (to go)
allons (we go)	→ allons	→ aller (to go)

Input	Stem	Lemma
genoux (knees)	→ genoux	→ genou (knee)
genou (knee)	→ genou	→ genou (knee)

Input	Stem	Lemma
travaux (roadworks)	→ traval	→ travail (work)
travail (work)	→ travail	→ travail (work)

4. GERMAN

4.1. Collisions

Identically spelled words with different meanings create the same stems, resulting in decreased *precision*.

Input	Stem	Lemma
Robbe (seal)	→ robb	→ Robbe (seal)
robbe (I crawl)	→ robb	→ robben (to crawl)

Input	Stem	Lemma
verdienst (you earn)	→ verdienst	→ verdienen (to earn)
Verdienst (income)	→ verdienst	→ Verdienst (income)

Input	Stem	Lemma
Frage (question)	→ frag	→ Frage (question)
frage (I ask)	→ frag	→ fragen (to ask)

Input	Stem	Lemma
spielen (they play)	→ spiel	→ spielen (to play)
Spielen (games)	→ spiel	→ Spiel (game)

Die grosse Robbe ist ins Wasser gesprungen. (The big seal jumped into the water.)
Ich robbe auf dem Boden um es zu erreichen. (I crawl on the floor in order to reach it.)

Du verdienst viel mehr als ich. (You earn much more than I do.)
Sein Verdienst vom letzten Jahr waren weniger als im Vorjahr. (His earnings from last year were less than the previous year.)

Ich habe eine Frage für dich. (I have a question for you.)
Ich frage ihn, ob er mit uns zusammenarbeiten will. (I ask him if he wants to work with us.)

Die Kinder spielen sehr gut zusammen. (The children play together very well.)
Sie hatten eine Menge Spaß mit den Spielen. (They had a lot of fun with the games.)

4.2. Collisions

Differently spelled words create the same stem, resulting in decreased *precision*.

Input	Stem	Lemma
schreiben (to write)	→ schreib	→ schreiben (to write)
Schreiber (scribe)	→ schreib	→ schreiber (scribe)

Input	Stem	Lemma
lesen (to read)	→ les	→ lesen (to read)
Leser (reader)	→ les	→ leser (reader)

Input	Stem	Lemma
Sau (sow)	→ sau	→ sau (sow)
sauer (sour)	→ sau	→ sauer (sour)

Input	Stem	Lemma
winden (to wind)	→ wind	→ winden (to wind)
Wind (the wind)	→ wind	→ wind (the wind)

Es ist sehr schwierig, mit der linken Hand zu schreiben. (It's very difficult to write with my left hand.)
Der Schreiber dokumentiert alles, was in der Sitzung gesagt wurde. (The scribe documented all that was said in the meeting.)

Meine Tochter lernte mit fünf Jahren lesen. (My daughter learned to read at age five.)
Der deutsche Leser hat einen starken Akzent. (The German reader has quite a strong accent.)

Die Sau gebar fünf Ferkel. (The sow gave birth to five piglets.)
Diese Milch ist ziemlich sauer. (This milk is quite sour.)

Ich muss ihm sagen, um den Faden wieder winden. (I must tell him to wind the thread again.)
Der Wind war so laut, dass es mich gestern Abend weckte. (The wind was so loud that it woke me up last night.)

4.3. Failure to Normalize

Words not normalized when stemmed result in decreased *recall*.

Input	Stem	Lemma
ging (went)	→ ging	→ gehen (to go)

Input	Stem	Lemma
schlief (slept)	→ schlief	→ schlafen (to sleep)

Input	Stem	Lemma
Onkeln (uncles)	→ onkeln	→ Onkel (uncle)

Input	Stem	Lemma
sah (saw)	→ sah	→ sehen (to see)

Ich ging früh nach Hause weil ich krank war. (I went home early because I was sick.)

Ich schlief bis zum Mittag. (I slept until noon.)

Ich gehe dort mit meinen drei Onkeln. (I go there with my three uncles.)

Ich sah ihn in den Aufzug. (I saw him in the elevator.)

4.4. Inconsistent Stems

Different stems for different inflections of the same word result in decreased *recall*.

Input	Stem	Lemma
trinken (we drink)	→ trink	→ trinken (to drink)
trinkst (you drink)	→ trinkst	→ trinken (to drink)
trinke (I drink)	→ trink	→ trinken (to drink)

Input	Stem	Lemma
arbeiten (we work)	→ arbeit	→ arbeiten (to work)
arbeitest (you work)	→ arbeitest	→ arbeiten (to work)

Input	Stem	Lemma
malen (we paint)	→ mal	→ malen (to paint)
malt (s/he paints)	→ malt	→ malen (to paint)

Input	Stem	Lemma
fahren (we drive)	→ fahr	→ fahren (to drive)
fährst (you drive)	→ fährst	→ fahren (to drive)

4.5. Compound Words

Stems do not decompound, resulting in decreased *recall*.

Input	Stem	Lemma
Küstenfischerei (coastal fisheries)	→ küstenfischerei	→ küste, fischerei

Input	Stem	Lemma
Telefonnummer (telephone number)	→ telefonnumm	→ Telefon, Nummer

Input	Stem	Lemma
Tischdecke (tablecloth)	→ tischdeck	→ tisch, decke

Input	Stem	Lemma
Lederjacke (leather jacket)	→ lederjack	→ leder, jacke

Er arbeitet mit der <u>Küstenfischerei</u> in der Nordsee. (He works with the coastal fisheries on the North Sea.)

Ich verlor ihre <u>Telefonnummer</u>, also ich schickte ihr eine E-Mail. (I lost her telephone number, so I sent her an email.)

Die <u>Tischdecke</u> ist voller Flecken. (The tablecloth is full of stains.)

Ich kaufte mir eine <u>Lederjacke</u> in Florenz. (I bought a leather jacket in Florence.)

5. ITALIAN

5.1. Collisions

Identically spelled words with different meanings create the same stems, resulting in decreased *precision*.

Input	Stem	Lemma
lavoro (work, job)	→ lavor	→ lavoro (work, job)
lavoro (I work)	→ lavor	→ lavorare (to work)

Input	Stem	Lemma
volo (flight)	→ vol	→ volo (flight)
volo (I fly)	→ vol	→ volare (to fly)

Input	Stem	Lemma
abiti (dresses)	→ abiti	→ abito (dress)
abiti (you live)	→ abiti	→ abitare (to live)

Input	Stem	Lemma
sorriso (a smile)	→ sorris	→ sorriso (a smile)
sorriso (smiled)	→ sorris	→ sorridere (to smile)

Ha fatto un lavoro eccellente. (He did an excellent job.)
Io lavoro in un ufficio in fondo alla strada. (I work in an office down the street.)

Il volo è arrivato tardi. (The flight arrived late.)
Io volo da Roma a Milano ogni settimana. (I fly from Rome to Milan every week.)

Ha comprato due abiti differenti per le nozze. (She bought two different dresses for the wedding.)
Tu abiti molto vicino al mio ristorante preferito. (You live very near my favorite restaurant.)

C'era un grande sorriso sul suo volto quando vide il clown. (There was a big smile on his face when he saw the clown.)
Ho sorriso quando ho sentito la notizia. (I smiled when I heard the news.)

5.2. Collisions

Differently spelled words create the same stem, resulting in decreased *precision*.

Input	Stem	Lemma
passato (the past)	→ passat	→ passato (the past)
passate (you spend)	→ passat	→ passare (to spend)

Input	Stem	Lemma
luna (moon)	→ luna	→ luna (moon)
lunar (lunar)	→ luna	→ lunar (lunar)

Input	Stem	Lemma
solo (only)	→ sol	→ solo (only)
sole (sun)	→ sol	→ sole (sun)

Input	Stem	Lemma
fine (end)	→ fin	→ fine (end)
fino (until)	→ fin	→ fino (until)

Si può imparare molto se si guarda al passato. (You can learn a lot if you look into the past.)
Passate tanto tempo davanti al computer. (You spend so much time in front of the computer.)

La luna è piena stasera. (The moon is full tonight.)
Il razzo ha fatto un atterraggio lunare. (The rocket made a lunar landing.)

Ho solo tre euro nel portafoglio. (I only have three Euros in my wallet.)
Il sole è nascosto dietro le nuvole. (The sun is hidden behind the clouds.)

Sono così stanco alla fine della giornata. (I am so tired at the end of the day.)
Io ti aspetterò fino alle 3. (I will wait for you until 3:00.)

5.3. Failure to Normalize

Words not normalized when stemmed result in decreased *recall*.

Input	Stem	Lemma
voci (voices)	→ voci	→ voce (voice)

Input	Stem	Lemma
mangi (you eat)	→ mangi	→ mangiare (to eat)

Input	Stem	Lemma
pomeriggi (afternoons)	→ pomeriggi	→ pomeriggio (afternoon)

Input	Stem	Lemma
guidi (you drive)	→ guidi	→ guidare (to drive)

Ho sentito voci nella stanza accanto. (I heard voices in the next room.)

Tu mangi troppa pasta. (You eat too much pasta.)

Sono alla scuola due pomeriggi a settimana. (I am at the school two afternoons per week.)

Guidi troppo veloce nella tua nuova Fiat. (You drive too fast in your new Fiat.)

5.4. Inconsistent Stems

Different stems for different inflections of the same word result in decreased *recall*.

Input	Stem	Lemma
ho (I have)	→ ho	→ avere (to have)
hanno (they have)	→ hann	→ avere (to have)
abbiamo (we have)	→ abbiam	→ avere (to have)

Input	Stem	Lemma
sono (I am)	→ son	→ essere (to be)
siamo (we are)	→ siam	→ essere (to be)

Input	Stem	Lemma
gioca (he/she plays)	→ gioc	→ giocare (to play)
giochiamo (we play)	→ giochiam	→ giocare (to play)

Input	Stem	Lemma
parlo (I speak)	→ parl	→ parlare (to speak)
parliamo (we speak)	→ parliam	→ parlare (to speak)

6. SPANISH

6.1. Collisions

Identically spelled words with different meanings create the same stems, resulting in decreased *precision*. According to our analysis, this occurs in 7% of Spanish word families.

Input	Stem	Lemma
prensa (media)	→ prens	→ prensa (media)
prensa (he/she presses)	→ prens	→ prensar (to press)

Input	Stem	Lemma
traje (dress)	→ traj	→ traje (dress)
traje (I brought)	→ traj	→ traer (to bring)

Input	Stem	Lemma
casa (house)	→ cas	→ casa (house)
casa (he/she marries)	→ cas	→ casar (to marry)

Input	Stem	Lemma
fue (he/she went)	→ fue	→ ir (to go)
fue (he/she was)	→ fue	→ ser (to be)

Me he comprado este hermoso traje para la fiesta. (I had bought this beautiful dress for the party.)
Yo traje cinco paquetes conmigo. (I brought five packages with me.)

Ella se casa mañana en la Iglesia de San Fermín. (She will marry tomorrow at the San Fermín church.)
Mi casa es de color blanco con ventanas rojas. (My house is white with red windows.)

Él fue a la piscina la semana pasada. (He went to the pool last week.)
Ayer fue un día muy aburrido. (Yesterday was a very boring day.)

6.2. Collisions

Differently spelled words create the same stem, resulting in decreased *precision*. According to our analysis, this occurs in 43% of Spanish word families.

Input	Stem	Lemma
publicaciones (publications)	→ public	→ publicación
público (public)	→ public	→ público

Input	Stem	Lemma
cómo (how)	→ com	→ cómo (how)
come (he/she eats)	→ com	→ comer (to eat)

Input	Stem	Lemma
qué (what)	→ que	→ qué (what)
que (than)	→ que	→ que (than)

Input	Stem	Lemma
solo (alone)	→ sol	→ solo (alone)
sólo (only)	→ sol	→ sólo (only)
sol (sun)	→ sol	→ sol (sun)

La prensa escrita se refiere a publicaciones impresas. (The written media refers to newspapers.)
Esta reunión no es pública. (This meeting is not public.)

¿Cómo estás? (How are you?)
No come carne casi nunca. (She hardly ever eats meat.)

No sé qué hacer. (I don't know what to do.)
Es más alto que su padre. (He is taller than his father.)

Lo haré yo sola. (I'll do it myself.)
Sólo quiero café. (I just want a coffee.)
Hace sol. (It’s sunny.)

6.3. Failure to Normalize

Words not normalized when stemmed result in decreased *recall*. According to our analysis, this occurs in 2% of Spanish word families.

Input	Stem	Lemma
lápices (pencils)	→ lapices	→ lápiz (pencil)

Input	Stem	Lemma
voces (voices)	→ voces	→ voz (voice)

Input	Stem	Lemma
duerme (he/she sleeps)	→ duerme	→ dormir (to sleep)

Input	Stem	Lemma
doy (I give)	→ doy	→ dar (to give)

Tengo dos lápices. (I have two pencils.)

Me gustan los voces. (I like the voices.)

Él duerme ocho horas diarias. (He sleeps 8 hours per day.)

Te doy mi palabra. (I give you my word.)

6.4. Inconsistent Stems

Different stems for different inflections of the same word result in decreased *recall*. According to our analysis, this occurs in 7% of Spanish word families.

Input	Stem	Lemma
he (I have)	→ he	→ haber (to have)
han (they have)	→ han	→ haber (to have)
hemos (we have)	→ hem	→ haber (to have)

Input	Stem	Lemma
pido (I ask for)	→ pid	→ pedir (to ask)
pedimos (we ask for)	→ ped	→ pedir (to ask)

Input	Stem	Lemma
caigo (I fall)	→ caigo	→ caer (to fall)
cae (he/she falls)	→ cae	→ caer (to fall)

Input	Stem	Lemma
hago (I do)	→ hago	→ hacer (to do)
hace (he/she does)	→ hace	→ hacer (to do)

7. CONCLUSION

As shown these examples, lemmatization boosts precision and recall by being language sensitive. In languages with compound words, decompounding allows individual word components to show up in search results, further boosting recall. In many cases, the search quality difference between using lemmatization vs. stemming can be quite significant.

8. EXPLORE FURTHER

For more information or to request an evaluation, please call us at 617-386-2090 or 800-697-2062, or write to info@basistech.com. We will be happy to assist you in evaluating the performance of our products on your data.