¿Que es Procesamiento de Texto Natural?

Un relato coloquial y crudo de mi comienzo en estas ciencias..
*by **Andrés Hohendahl** as of 2nd Qtr. 2009*

Hace unos años.. (por el 2005) decidí desarrollar (en realidad intentar) un sistema de conversación por SMS en castellano y puse manos a la obra…
(no sabía en lo que me había metido!)

Estudié el tema un poco más por lo que me puse a tratar de conseguir un diccionario español..

Lo cual me resultó bastante difícil (en especial por el tema licencias y derechos).

Los diccionarios que conseguí gratis eran espantosos, por su pobre cobertura de palabras y mala calidad, ni hablar que decían poco y nada de las palabras halladas (eran básicamente para corrección ortográfica *spell-correct*) Luego me tomé con los oficiales Espasa, Encarta, Salvat, Larousse, María Moliner, (Locademias de letras) como Academia Argentina de Letras, y los más serios tipo Real Academia Española. Todo ellos tienen definiciones no-normalizadas, es decir entras con una palabra y... si tenés suerte.. salís con un montón de palabras... abreviaturas propias, en un orden poco ortodoxo, pero en general no tenés la mas reputa idea de que es cada una, eso si si las metes en una base dedatos y las buscas como resultado… genial.. pero si las volvés a buscar a todas, obtenés otro montón de palabras más grande aún (uno por cada palabra buscada.. y así) pero para obtener algo útil para procesar, eso… es otro cantar. Para colmo de estos diccionarios, no hay versiones para desarrolladores, tampoco te dan la base de datos ni por casualidad! Traté de contactar la Real Academia española, la Argentina y caso todas las editoriales como Larousse, y no me dieron una respuesta potable.

Ah! y los investigadores… ellos tienen (en países mas avanzados) acceso a algunos diccionarios, pero la mayoría prueban sus algoritmos sobre un 'set' o conjunto de pruebas que ellos mismos arman, por lo general acotados; los cuales son geniales para sus desafíos, pero no sirven de nada para un sistema real que interactúe con la gente.

Entrando en tema más profundamente, descubrí que aquellos diccionarios convencionales, por más que pueda transcribirlos, además no sirven de mucho, pues solamente tienen las palabras raíz (llamado lema). Es decir verbos están en infinitivo (amar/correr/decir) y los sustantivos y adjetivos en singular y masculino (salvo vaca). No contienen peyorativos, diminutivos, aumentativos y demás accidentes morfológicos.

Eso es muy malo (pensé) en otras palabras no te sirve un diccionario por más que lo consigas o lo armes. Además la tarea megalómana de armar un diccionario con los verbos conjugados es monumental, por no decir imposible para una pequeña organización (son mas de 200.000 horas-lexicógrafo), para que lo puedas evaluar simplemente, estos son los datos: 35.000 verbos x 2 géneros x 3 personas (yo-tu-el) x 2 plural/singular x 6 tiempos (presente/pasado/indicativo/subjuntivo/pretéritos/gerundios/participios/participios activos) eso da como 70 variaciones.

Si fuese fácil genial! lo hago a mano.. pero hete aquí las malas noticias: los verbos españoles tienen mas de 49 maneras diferentes de conjugarse!, y NADIE te dice a cual manera pertenece cada verbo de los 30 mil y pico. *(ni el diccionario de verbos y conjugaciones de Larousse cubre al de la Real Academia)* . A eso hay que pensar lo siguiente : por cada verbo/ sustantivo/ adjetivo/ adverbio hay variaciones por prefijos: *de- pre- auto- ante- post- in-* (y muchos mas,,, suman mas de 900 en general) es decir: 900 x 35.000 x 70 es un número muy pero muy grande. Resumiendo un diccionario con capacidad morfológica debe poder reconocer al menos 2-5 millones de palabras, idealmente deberían ser como 30 millones basado en 50 a 80 mil raíces comunes.

Eso es monumental hacerlo y encima eficientemente, además se puede comprar hecho.. genial! ¿dónde? (en Ecuador y/o Finlandia) en donde saben menos Español que nosotros y estimo lo armaron automáticamente por traducción (CONEXXOR). La licencia? ah! (pequeño detalle) es por cliente/usuario/tipo_de_aplicación/servidor/etc. (mas compleja que las de microsoft) y bastante caras por cierto (7000 euros para arriba, hasta 50.000 euros anuales para una licencia "open" para usar con celulares y algunos Millones de usuarios en un solo país, que era mi caso) Les pedí una licencia de DEMO (Académica por la UBA) para probar el producto, un desastre! un DLL en windows, de 12 megas, complicada de instalar, no ampliable, tarda un huevo en cargar, es leeeeeenta y le pifia a muchas de las palabras mas comunes!, no tiene listado ni detalle de contenidos y alcances, no es customizable, las etiquetas que te entrega no están normalizadas (salvo por ellos) y la documentación es horrorosa... mente pobre y confusa, en resumidas: es una “CAJA NEGRA” bastante cara y sin muchas garantías.

Esto, sumando al precio me decidió ponerme a estudiar el tema y buscar soluciones viables..

No fué sencillo y decidí volver a mi época de las cruzadas... como investigador de la UBA y Conicet, me fui a la UBA, consulté a mis antiguos compañeras (hoy docentes) pues yo fui investigador como por 10 años, ostento el premio nacional de electrónica del 1985.
Me re-inscribí en la UBA, y cursé + aprobé 10 materias optativas y algunas para doctorado, para actualizarme luego de 20 años de no pisar la UBA, y de paso completar los créditos para poder presentar una tesis y hacer un doctorado en ingeniería. Luego de 6 meses de cursar entre Exactas y la UBA + haciendo contactos con todo el mundo académico... me encontré este panorama: *no hay un carajo de nadie que sepa suficiente de esto en Argentina y menos que esté dispuesto a compartir conocimiento y/o trabajo*, hay unos pocos en España, México y casi nadie en USA (*con todas las universidades de reputa-madre como: Harvard, MIT, UCSD, UPENN, CMU, etc.*) Eso si en inglés tienen algo más (tampoco gratis ni potable)

¡Que panorama!
Ante esto.. la única solución era hacerlo yo… o dedicarme a otra cosa!

Y como no me suelo dejar vencer… Puse manos a la obra.. me replegué, invertí tiempo, leí, consulté a lingüistas, estudié lingüística.. trabajé mucho y desarrollé (inventé) algunas cosas nuevas. Cuando me parecieron potables las publiqué para someterlas a juicio académico y/o criterio del estado del arte.. y me las aceptaron en algunos congresos internacionales (CACIC 2006) y revistas. Luego presenté más papers (WICC 2006) y me fui armando de un paquete de utilidades para el español e inglés, que hoy me permite decir que debo tener uno de los únicos lematizadores EFICIENTES y PROTABLES (con código fuente en C# de autoría propia) capaces de atacar problemas tan complejos como lo es el castellano de hoy... además con éxito. Luego y como me gustó… fui por más!

Hice el primer diccionario-traductor con capacidad de reconocer verbos conjugados, palabras el plural, diminutivo y demás. Lo puse a funcionar para SMS y lo ofrecí al estado (Ministerio de Educación) ¿El resultado? Ya saben: no les interesa aportar a la cultura nacional (*a ver si se avivan y no los votan mas*).. pero como no me dejé vencer.. se lo ofrecí a un país al cual le interesa su cultura: Ecuador y se lo vendí a MoviStar, está funcionando desde el 2005.

Con esta experiencia, me encontré con los errores de la gente consultando por SMS con las variantes de palabras inexistentes, al igual que en el chat.. (*eso es otro kilombo mayúsculo.*)

Ud. quería decir…
Construí luego un sistema de corrección de errores.. pero.. aquí aparecieron otras cuestiones no simples ni fáciles de abordar: ¿como decidir cual es la palabra correcta? si tengo como 3 o 5 similares, ¿cual es el criterio? similitud... como ¿letras iguales? ¿por como suena? y cómo era eso de las reglas de ortografía.. todo un tema!.

Me suena que me dijo…
Investigué sobre similitud fonética, y no había nada en el horizonte científico sobre el tema, solo un pobre algoritmo en desuso llamado Soundex y uno llamado Metaphone (pero son índices fonéticos, no miden similitud), para colmo andan para el inglés, mas o menos, pero para el español.. ¡bien gracias! Ni por puta!

Y entonces.. decidí inventar uno: (como trabajé varios años en investigación aplicado a fonoaudiología y fui co-autor del primer implante intra-coclear en 1982 en sudamérica) Decidí inventar una teoría nueva basada en un modelo fono-articulatorio.. lo armé y lo probamos... con mas de 100 personas reales. El resultado? Una grata sorpresa! Más del 85% de coincidencia con lo que piensa un humano de cuanto se parecen las palabras.. lo sometimos a escrutinio académico y lo presentamos en el "*X Congreso Latinoamericano de Neuro Psicología*" (SLAN 2007) todo una sensación, NADIE había hecho algo similar. Fuimos los primeros en el Mundo! (lo cual no dice mucho, pero es una satisfacción)

Ahora el kilombo es.. tengo las palabras, sé que quieren decir, tengo sus raíces, sus múltiples funciones gramaticales y significados,.. ahora ¿cual de ellos es el que va en esta conversación?

Eso te lleva al tercer desafío:

LA DESAMBIGUACIÖN
No es tema fácil, en el mundo (*estado del arte*) lo hacen estadísticamente, con unos métodos bastante complejos (*Markov, SVM, Maximum Entropy, Simulated Anhealing, Relaxation, etc.*) todos ellos geniales en su tipo.. pero pocos funcionan en castellano y necesitan un diccionario lematizador como mínimo. Tampoco andan bien con habla espontánea[1]. Hay que entrenarlos con un *corpus* y no logran demasiado éxito (*tienen un techo acotado*).

Una vez que (supongamos) tenemos todo funcionando bien, sabemos que (por ejemplo):

"EL GATO COME CARNE HOY"

"***artículo-sustantivo-verbo-sustantivo-adverbio***"

Genial! Pero entonces ¿que cuernos hacemos con esto? ¿cuál palabra consigna la persona que habla? ¿cuál es el sujeto, el predicado, el objeto directo, el circunstancial? ¿dónde esta?..

Si, lamentablemente me recuerda ***eso tan HORRIBLE de castellano de la secundaria***
Y por último los sentidos de cada significado: GATO se refiere a una persona con habilidades felinas, una mina-puta o un animal-felino?
Esto por lo general se lo llama WSD *(Word-Sense-Desambiguation)* y hay bastante poco en este tema que ande decentemente y sea robusto[1]. Luego hay que entender eso bajo un esquema de lenguaje (*gramática española*) eso de saber cual e cada parte, se llama en espanglish: *parseo* (***parsing***), pero el drama es que esta gramática nuestra.. es muy ambigua y compleja. Los sistemas convencionales de ***parsing*** por lo general fallan en su etapa más temprana y teórica, sin siquiera poder abordar el problema (LR / LALR / LLk) Hay un único tipo que sirve un poco, y es el GLR, pero tampoco hay mucho hecho con esto.. y menos para el castellano y accesible.

¿Qué dijo?
Luego.. y para poder entender que demonios dijo el usuario.. hay que hacer algo para entender lo que quiso decir el usuario en español (*Deep Understanding*)… **Ese tema está hoy abierto!**
Y no hay mucho que funcione ni siquiera potablemente[1], unos pocos con vocabularios muy acotados en inglés, chino, ruso y alemán… pero como por supuesto… casi nada en Español.

Y para contestar en forma coloquial?
Bueno podríamos poner todas las respuestas ya escritas, suponiendo que sabemos todo lo que nos van a preguntar, pero como eso es imposible…ni permitiría que un sistema "*elabore una respuesta*" entonces hay que poder elaborar una frase en base a conocimiento de algo:
Por ejemplo poder elegir un verbo, luego conjugarlo en el tiempo, modo, persona y número apropiados, ponerle un artículo con el mismo género y número del sustantivo al cual afecta, elegir el adjetivo que le corresponde, ponerlo en diminutivo con el género y número apropiado, poner el circunstancial.. y… decirlo!

¿con qué se hace todo esto?
Rta.: con un montón de trabajo, horas-culo, ganas e ingenio.. y creo estar en el camino!

¡Bienvenidos al Fascinante Mundo del Procesamiento de Lenguaje Natural!

Andrés Hohendahl

[1]* *Cuando digo algo que funcione, hablo de algo útil y aplicable a la vida real.*
No hablo de un "super-paper-académico" presentado en un congreso de reputa-madre que arrojó en un F-Score del 82%, con un grupo acotado de frases con un mini-diccionario ajustado inteligentemente al experimento para que funcione.