Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages

Ezeiza N., Alegria I., Arriola J.M., Urizar R.
Informatika Fakultatea
649 P.K Donostia E-20080
jibecran@si.ehu.es
http://ixa.si.ehu.es

Aduriz I.
UZEI
Aldapeta, 20.
Donostia E-20009
uzei@sarenet.es

Laburpena

Artikulu honetan metodo estokastiko eta erregeletan oinarritutako metodoen arteko konbinaketa euskarari aplikatzearen emaitzak aurkeztuko ditugu.Desanbiguazioan erabilitako metodoak Murrizpen Gramatika (CG) eta MULTEXT proiektuak garatutako HMMn oinarritutako etiketatzailea dira.
Euskara hizkuntza eranskaria izaki, hitz bakoitzari dagozkion irakurketa guztiak esleitzeko analizatzaile morfologikoa beharrezkoa da. Ondoren, CG erregelak informazio morfologiko guztiari aplikatzen zaizkio eta prozesu honek testuen anbiguotasuna gutxitzen du. Azkenik, geratutako etiketen artean bakarra hautatzeko MULTEXT proiektuko tresnak erabiltzen dira.
Metodo estokastikoa soilik erabiltzean, errore-tasa %14 ingurukoa da, baina etiketatzailearen doitasuna hitz ezezagunekin lexikoa aberastuz gero %2 hobe daitekeen arren. Metodo biak konbinatzen direnean, berriz, prozesu osoaren errore-tasa %3.5ekoa da. Ikasketarako corpusa nahikoa txikia dela, HMM eredua lehenengo mailakoa eta euskararako Murrizpen Gramatika oraindik ere garapen prozesuan dagoela kontuan izanik, gure ustez metodo konbinatu hau erabilita emaitza onak lor daitezke eta beste hizkuntza eranskarietarako bereziki egokia izan daiteke.

Resum

En aquest article presentem els resultats de la combinació de mètodes estocàstics i basats en regles aplicats a la desambiguació morfosintàctica de l'euskara. Els mètodes utilitzats per a la desambiguació són: les Gramàtiques de Restriccions (CG) i l'etiquetador basat en HMM del projecte MULTEXT.
El caràcter aglutinant de l'euskara fa necessari la utilització d'un analitzador morfològic per assignar a cada paraula totes les seves interpretacions. Les regles de CG s'apliquen utilitzant la informació morfològica completa i aquest procés redueix parcialment l'ambigüitat dels textos. A continuació, s'apliquen les eines de MULTEXT per escollir una única etiqueta.
Utilitzant només el mètode estocàstic la taxa d'error és aproximadament del 14%, encara que la precisió de l'etiquetador es pot incrementar en un 2% utilitzant les paraules desconegudes per enriquir el lèxic. En canvi, la combinació d'ambdós mètodes permet reduir l'error fins al 3.5%.
Tenint en compte que el corpus d'aprenentatge és bastant petit, que el model HMM és de primer ordre i que la Gramàtica de Restriccions de l'euskara està encara en fase de desenvolupament, creiem que els resultats del mètode combinat són bons i que la combinació de mètodes és especialment adequada per a llengües aglutinants.

Resumen

En este artículo presentamos los resultados de la combinación de métodos estocásticos y basados en reglas aplicados al euskara. Los métodos utilizados para la desambiguación son las Gramáticas de Restricciones (CG) y el etiquetador basado en HMM del proyecto MULTEXT.
Siendo el euskara una lengua aglutinante, será necesario un analizador morfológico para asignar a cada palabra todas sus interpretaciones. A continuación se aplican las reglas de CG utilizando toda la información morfológica y este proceso disminuye la ambigüedad de los textos. Por último, las herramientas de MULTEXT escogerán una única etiqueta.
Utilizando únicamente el método estocástico la tasa de error es de alrededor del 14%, aunque la precisión del etiquetador puede incrementarse en un 2% utilizando las palabras desconocidas para enriquecer el léxico. En cambio, combinando ambos métodos la tasa de error del proceso completo es del 3.5%. Teniendo en cuenta que el corpus de aprendizaje es bastante pequeño, que el modelo HMM es de primer orden y que la Gramática de Restricción del euskara está aún en fase de desarrollo, creemos el método combinado obtiene buenos resultados y puede ser adecuado para otras lenguas aglutinantes.