Videókhoz kapcsolódó kiegészítő információk többnyelvű keresése a Wikipédia segítségével

Gyarmati Ágnes[1], Gareth J.F. Jones[1]

[1] Dublin City University, Centre for Digital Video Processing,
Dublin 9, agyarmati@computing.dcu.ie

Kivonat: (ld. a törzsben)[1]

1 Rövid kivonat

Az egyre nagyobb mennyiségben szabadon elérhető digitális tartalmak (hang- és videófelvételek) akkor válnak igazán értékessé, ha a felhasználók könnyen megtalálhatják a számukra érdekes tartalmakat, részleteket, azaz hatékonyan lehet köztük és bennük keresni.

Az 1990-es évek közepén a TREC kutatói fórum egyik céljaként tűzte ki a hangzó szövegben való keresés hatékonyságának fejlesztését. Ehhez amerikai rádiós híradásokat használtak, és csupán pár év elteltével a próbálkozásokat sikeresnek, a problémát gyakorlatilag megoldottnak tekintették [1].

Később azonban felmerült az igény, hogy a viszonylag zárt szókincsű, gondozott beszéddel felolvasott szöveget rögzítő jó minőségű stúdiófelvételeken túl az élőbeszéddel is érdemben foglalkozzanak az információ-visszakeresés területén. Erre például a CLEF fórum beszéd-visszakereső modulja vállalkozott [3].

A VideoCLEF 2009 kiírása más szemszögből nyújtott vizsgálati lehetőséget élőbeszéd és keresés kapcsolatához [2]. Az úgynevezett „Összekapcsoló feladat" (Linking Task) egy kulturális dokumentumműsor egyes rövid, csupán pár másodperces részleteihez keres tartalmilag kapcsolódó oldalakat a Wikipédiában. A feladat nehézségét a híranyagoknál tapasztaltaknál spontánabb beszéd és változatosabb tartalom mellett egy nyelvi csavar adja: a dokumentumsorozat holland nyelvű, míg a linkelésnél az angol Wikipédiából kellett keresni a legmegfelelőbb oldalakat – ezáltal szimulálva egy haladó nyelvtanulót, aki már rendelkezik kellő nyelvismerettel, hogy idegen nyelven is érdemes legyen tévét, videót, műsorokat néznie, de még szükségét érezheti, hogy a számára érdekes vagy magyarázó kiegészítő információkhoz anyanyelvén, vagy legalábbis egy általa magasabb szinten beszélt (értett) harmadik nyelven szeretne hozzájutni (esetünkben angolul).

Alapvetően két különböző megközelítés kínálkozik. Az egyik a holland Wikipédiát veszi alapul, ott végzi a keresést az eredeti holland szöveg felhasználásával, majd a

[1] Ez a kutatás a Science Foundation Ireland (SFI) által támogatott Improving Indexing for Search of Spontaneous Conversational Speech (IISSCoS) projekt keretében zajlik.

Wikipédia saját, nyelvek közötti linkjeit követve jut el a relevánsnak tartott angol oldalakig. A másik előbb gépi fordítással angol nyelvű szöveget gyárt a műsorok holland átirataiból, majd közvetlenül az angol Wikipédiában keresi a megfelelő oldalakat. Az előadás e két módszert kívánja tárgyalni, bemutatni előnyeiket, hátrányaikat, különféle változataikat, melyek pl. az adatok használatában, keresőkifejezések generálásában is különbözhetnek. Eddigi eredményeink azt mutatják, hogy a keresés szempontjából nem hatékony az átiratot automatikusan lefordítani, hanem inkább a forrásnyelven érdemes a keresést végezni.

A felvázolt módszerek bármely nyelvre, nyelvpárra alkalmazhatók, így a magyarra is, akár tárgy-, akár célnyelvként, feltéve, hogy létezik az adott (a videóban használt nyelvre) automatikus beszédfelismerő program, és hogy a kívánt nyelvű Wikipédia kellően gazdag.

Bibliográfia

1. John S. Garofolo, Cedric G. P. Auzanne, and Ellen M. Voorhees. The TREC spoken document retrieval track: A success story. In Text Retrieval Conference (TREC) 8, 16–19, (2000).
2. M. Larson, E. Newman and G. J. F. Jones, Overview of VideoCLEF 2009: New Perspectives on Speech-based Multimedia Content Enrichment, In Proceedings of the CLEF 2009: Workshop on Cross-Language Information Retrieval and Evaluation, Korfu, Görögország, (2009). Megjelenés alatt.
3. D. W. Oard, J. Wang, G. J. F. Jones, R. W. White, P. Pecina, D. Soergel, X. Huang and I. Shafran, Overview of the CLEF-2006 Cross-Language Speech Retrieval Track, In Proceedings of the CLEF 2006: Workshop on Cross-Language Information Retrieval and Evaluation, Alicante, Spanyolország, Springer (2006)