Critères numériques dans les essais cliniques : annotation, détection et normalisation

Natalia Grabar[1] Vincent Claveau[2]
(1) CNRS, Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
(2) CNRS, IRISA, Rennes, France
natalia.grabar@univ-lille3.fr, vincent.claveau@irisa.fr

RÉSUMÉ
Les essais cliniques sont un élément fondamental pour l'évaluation de nouvelles thérapies ou techniques de diagnostic, de leur sécurité et efficacité. Ils exigent d'avoir un échantillon convenable de la population. Le défi consiste alors à recruter le nombre suffisant de participants avec des caractéristiques similaires pour garantir que les résultats des essais sont bien contrôlés et dus aux facteurs étudiés. C'est une tâche difficile, effectuée essentiellement manuellement. Comme les valeurs numériques sont une information très fréquente et importante, nous proposons un système automatique qui vise leur extraction et normalisation.

ABSTRACT
Numerical criteria in clinical trials : annotation, detection and normalization.

Clinical trials are very important for the evaluation of new therapies or diagnosis procedures, of their security and efficency. They must rely on convenient sample of the population. the challenge is then to recruit sufficient number of participants with similar characteristics to guarantee that the results are controlled and due tu the factors studied. This is a difficult task, mainly performed manuelly. Because numerical values are frequent and provide important information, we propose an automatic system that aims their extraction and normalization.

MOTS-CLÉS : Valeurs numériques, extraction d'information, normalisation, domaine médical.

KEYWORDS: Numerical values, information extraction, normalization, medical area.

1 Introduction

Les essais cliniques sont un élément fondamental pour l'évaluation de nouvelles thérapies ou techniques de diagnostic, de leur sécurité et efficacité. Ils se basent sur l'inférence statistique et exigent d'avoir un échantillon convenable de la population. Le défi consiste donc à recruter un nombre suffisant de participants avec des caractéristiques spécifiques pour garantir que les résultats des essais sont bien contrôlés et dus aux facteurs étudiés et non pas à d'autres facteurs ou au hasard. Pour cette raison, les essais cliniques doivent définir précisément la population d'intérêt en indiquant les critères d'inclusion et d'exclusion. Ce sont par exemple l'âge, le genre, l'histoire médicale, les traitements, les biomarqueurs, etc. La source principale d'information sur les patients est le dossier patient, qui peut être en version papier ou électronique. Ces dossiers contiennent essentiellement du texte non structuré en langue naturelle. Pour cette raison, ce sont surtout les opérateurs humains (ie, investigateurs principaux et assistants de recherche clinique) qui sont capables de détecter efficacement

les patients éligibles (Campillo-Gimenez *et al.*, 2015). Il va sans dire que c'est une tâche laborieuse et longue, qui constitue un écueil dans le recrutement des patients. Il est en effet commun que les personnes qui mènent les essais cliniques n'arrivent pas à recruter les patients à temps et en un coût raisonnable (Fletcher *et al.*, 2012). Ainsi, à cause de la complexité croissante des protocoles et du processus de recrutement, presque la moitié de retards dans les essais cliniques sont dus à la difficulté de recrutement, alors que le pourcentage des essais qui terminent les recrutement à temps est très bas : 18 % en Europe, 17 % en Asie, 15 % en Amérique Latine, 7 % aux Etats-Unis (Center Watch, 2013). S'il existe des travaux qui proposent de développer les systèmes de recrutement électronique (Cuggia *et al.*, 2011), il reste difficile de générer une représentation comparable des informations contenues dans le texte libre des dossiers cliniques et les critères d'éligibilité (Embi *et al.*, 2005; Olasov & Sim, 2006; Ross *et al.*, 2010; Tu *et al.*, 2011; Pressler *et al.*, 2012; Shivade *et al.*, 2014). Pour cette raison, ces systèmes exploitent essentiellement les données structurées, en laissant de côté le texte narratif, et reposent ensuite sur l'expertise manuelle. Nous pensons que les méthodes de TAL peuvent être efficacement exploitées pour cette tâche. Nous proposons de nous concentrer sur les informations numériques ou quantifiées, qui sont fréquentes et importantes pour le recrutement des patients.

Il existe quelques travaux sur l'extraction des informations numériques à partir des données non structurées. Dans la langue générale, comme les textes de la Toile, un système pour l'extraction et approximation de valeurs numériques (comme la hauteur ou le poids) a été proposé (Davidov & Rappaport, 2010). Il repose sur des patrons de relation et WordNet. Les termes similaires sont d'abord obtenus grâce à WordNet et les valeurs numériques liées sont extraites. Ensuite, les objets extraits sont comparés. Ce système montre une précision moyenne exacte de 0,84 et inexacte de 0,72. Un autre travail propose deux systèmes d'extraction qui demande toutefois une supervision manuelle (Madaan *et al.*, 2016). Différents cas sont pris en charge, comme la durée de vie, l'inflation, la production d'électricité. Plus proche de nos besoins, quelques études sont à noter dans le domaine médical. A partir d'un petit ensemble de données cliniques annotées en français, un modèle CRF est entraîné pour reconnaître trois types d'entités (concepts, valeurs, unités). Ensuite, un système à base de règles est créé pour calculer les relations entre ces entités (Bigeard *et al.*, 2015). Des travaux similaires ont aussi été proposés pour l'anglais sur des données cliniques (Sarath *et al.*, 2016) et sur des rapports de radiologie cardiaque (Nath *et al.*, 2016).

Ces différents travaux obtiennent des résultats difficilement comparables entre eux du fait des différences de nature des textes traités, des informations recherchées et des protocoles d'évaluation. Il nous a ainsi semblé nécessaire de développer un jeu de données qui soit spécifiquement dédié à notre cas d'usage final (section 2). Nous nous intéressons ainsi aux protocoles des essais cliniques en anglais. Ils comportent en effet une très grande variété d'expressions numériques et de critères quantifiés. En plus de l'extraction, pour laquelle nous reprenons des techniques similaires à l'état de l'art (section 3), nous proposons une méthode pour aider à la normalisation des unités (section 4).

2 Données et annotations

2.1 Essais cliniques

211 438 protocoles d'essais cliniques, en anglais, ont été collectés (en décembre 2016 à partir du site `www.clinicaltrials.gov`). Parmi les différentes sections composant ces protocoles, nous nous intéressons uniquement aux critères d'inclusion et d'exclusion, ce qui constitue un corpus de

κ/phase	A1 vs. A2	A1 vs. A3	A2 vs. A3	A1 vs. A4
phase 1 global	0.78	0.51	0.47	-
détails (C, U, V)	0.72, 0.58, 0.83	0.34, 0.58, 0.57	0.22, 0.56, 0.7	
phase 2 global	-	-	-	0.92
détails (Q, T)				0.84 0.93

TABLE 1 – Accord inter-annotateur (κ de Cohen) sur les deux phases d'annotation

plus de 2 millions de phrases. Trois phrases tirées de ce corpus sont données ci-après :

— *Absolute neutrophil count $\geq$ 1,000 cells/μl at time of enrollment.*
— *Exclude if T3 uptake is less than 19% ; T4 less than 2.9 ((g/dL) ; free T4 index is less than 0.8.*
— *Elevated bilirubin within the past two years.*

2.2 Annotations de référence

1 500 phrases ont été sélectionnées aléatoirement et annotées chacune par plusieurs annotateurs. L'annotation a eu lieu en deux temps. Dans un premier temps, trois annotateurs (A1, A2, A3), avec des profils différents, ont annoté les phrases en repérant les valeurs (V) et unités (U) exprimées, ainsi que les concepts (C) auquel ces mesures se rapportent. Dans un second temps, deux annotateurs (A1, A4) ont annotés ces mêmes phrases en repérant les critères exprimés avec des quantificateurs (Q) non numériques (*normal, low, clinically important...*), ainsi que les expressions de validité temporelle (T). A2 est médecin, A1, A3, A4 sont informaticiens ; A1 et A2 ont une bonne expérience des données cliniques.

Ainsi, dans les exemples précédents, les annotations seraient :

— C : *Absolute neutrophil count, T3 uptake, T4, free T4 index, bilirubin* ;
— V : *$\geq$ 1,000, less than 19, less than 2.9, less than 0.8* ;
— U : *cells/μl, %, g/dL* ;
— Q : *Elevated* ;
— T : *at time of enrollement, within the past two years*

Pour évaluer l'accord inter-annotateur, nous utilisons la mesure du κ de Cohen (Artstein & Poesio, 2008) pour chaque paire d'annotateurs. Le résultats sont reportés dans le tableau 1 pour chacune des phases d'annotation (C, V, U d'une part et Q, T d'autre part). de manière globale et en détail pour chacun des types d'annotations. Ils montrent un très fort agrément sur la phase 2 entre les deux annotateurs. Concernant la phase 1, les résultats sont plus contrastés : A1 et A2 obtiennent le plus fort agrément, ce qui souligne l'intérêt d'une connaissance du domaine médical, en tant que médecin ou informaticien. Un examen des cas de désaccord montre qu'une grande part est due à l'annotation des concepts sur lesquels portent les mesures, notamment pour les accords impliquant A3. Les valeurs numériques sont en revanche bien reconnues, quelle que soit la formation initiale de l'annotateur.

Sur la base des annotations de A1, A2 et A4, une annotation finale est obtenue par discussion des désaccords jusqu'à l'obtention d'un consensus. Ces données annotées et le guide d'annotation sont accessibles à `people.irisa.fr/Vincent.Claveau/Corpus`.

3 Annotation automatique

Dans cette section, nous décrivons l'utilisation du jeu de données pour entraîner des modèles CRF pour détecter, dans de nouveaux textes cliniques, des concepts mesurés (C), soit précisément (avec valeur V et unité U), soit avec un quantificateur (Q), et le temps associés à cette mesure (T).

3.1 Méthodologie

Pour produire ces modèles d'annotation automatique, nous utilisons les champs aléatoires conditionnels (Conditional Random Fields, CRF) (Lafferty *et al.*, 2001). Ces modèles graphiques non dirigés permettant d'apprendre les distributions de probabilités d'annotations y sachant les observations x. Très populaires en TAL par leur capacité à prendre en compte l'aspect séquentiel du texte et à tirer parti de descriptions riches du texte, ils sont des outils standard en extraction d'information, reconnaissance d'entités nommées, étiquetage, etc. (Wang *et al.*, 2006; Pranjal *et al.*, 2006; Constant *et al.*, 2011; Raymond & Fayolle, 2010, inter alia).

Dans notre cas, nous adoptons un schéma d'étiquetage BIO : les CRF doivent assigner une étiquette à chaque mot selon qu'il dénote les éléments recherchés (C, V, U, T, Q) ou non (O) ; ces éléments pouvant être composés de plusieurs mots, nous adoptons le schéma d'annotation BIO (par exemple, B-C indique le premier mot d'un concept, I-C les mots suivants de ce concept, O pour un mot n'appartenant à aucune des cinq informations recherchées).

Les textes dont décrits par les mots-formes, les lemmes, les parties du discours (ces informations et la tokénisation sont obtenues par Treetagger (Schmid, 1994)), des indices graphémiques (présence de majuscules, de chiffres, de symboles...). Nous autorisons les CRF à regarder un contexte de 4 mots avant et après le mot courant. L'implémentation CRF que nous utilisons est Wapiti (Lavergne *et al.*, 2010). Pour mesurer les performances de cette tâche d'annotation automatique, nous évaluons les résultats obtenus sur une partie du jeu annoté manuellement (cf. section 2.2) en termes de taux d'erreur mot (pourcentage de mots recevant une étiquette différente de celle du jeu de référence) et de F-mesure (Sebastiani, 2002).

3.2 Résultats

La figure 1 présente les résultats en taux d'erreurs mots et la F-mesure de chacune des catégories (concept, valeur, unité, quantificateur, temps) en fonction de la quantité de données d'apprentissage utilisée. Pour chaque taille de jeu d'entraînement, le phrases d'entraînement sont tirées au hasard, le reste des données annotées sert de jeu de test, et ce processus est répété cinq fois puis les résultats moyennés pour éviter tout effet lié à la variabilité des petits jeux d'entraînement. Le taux d'erreurs global décroît continuellement, ce qui suggère que des données d'apprentissage supplémentaires permettraient d'améliorer encore les résultats. L'analyse des résultats par catégorie permet de comprendre que ce sont les concepts et les quantificateurs qui bénéficieraient le plus de ces données supplémentaires puisque leur rappel est bas et la détection des autres catégories atteint un plateau.

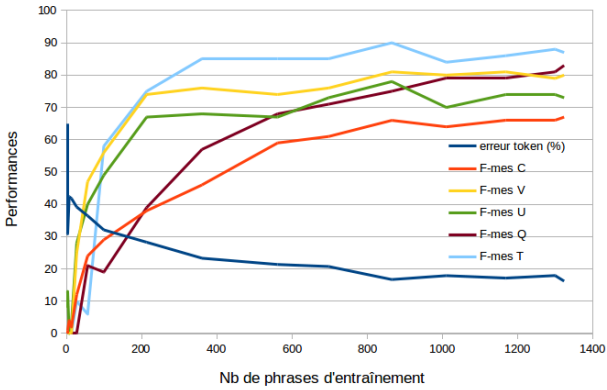

FIGURE 1 – Performance de l'annotation par CRF (globalement en terme de taux d'erreur mot, et par catégorie en terme de F-mesure) selon la quantité de données d'entraînement (nombre de phrases).

4 Vers une normalisation des unités

Une fois que les critères d'éligibilité recherchés sont détectés, il est nécessaire de les normaliser pour pouvoir les utiliser pour sélectionner des patients. C'est notamment le cas pour les unités qui apparaissent sous de nombreuses variantes :

— variation typographique (*cells/mL* vs. *cells / ml*) ;
— flexion (*cell/mL* vs. *cells/mL*) ;
— acronymie (*xULN* vs. *time Upper Limit of Normal*) ;
— variation paradigmatique (*cigarettes per day* vs. *packs per week*).

4.1 Détection des variations

Nous proposons une approche pour identifier, pour une expression d'unité donnée, toutes ses variations possibles et les ordonner par similarité. L'idée est d'utiliser les concepts associés à ces unités avec des techniques issues de la recherche d'information. Plus précisément, notre approche est la suivante. Tout d'abord, un modèle CRF, appris sur les données de référence, est utilisé pour annoter l'ensemble du corpus d'essais cliniques (i.e. les 2 millions phrases des critères d'éligibilité). Ensuite, chaque unité détectée est collectée ; celles avec exactement la même graphie sont groupées. Les concepts détectés avec ces unités sont décomposés en n-grammes et pondérés (TF-IDF). Chaque graphie d'unité est ainsi représentée par un vecteur de n-grammes (de concepts). La proximité entre deux graphies d'unités est alors définie par la proximité entre leurs vecteurs, et mesurée par un cosinus.

4.2 Résultats

Vingt graphies d'unités ont été tirées au hasard parmi celles détectées par CRF sur les essais cliniques. Pour chacune, nous évaluons la précision de la liste de ses n plus proches voisins (ppv) selon la similarité cosinus. Deux types de précision sont données : la précision 'unité identique' n'accepte

	P@1 (%)	P@10 (%)	P@1_freq (%)
unité identique	50	45	45
unité équivalente	100	84.5	95

TABLE 2 – Précisions des plus-proches voisins (moyenne sur 20 expressions d'unités choisies aléatoirement) au rang 1 et 10, et précision sur le voisin le plus fréquent parmi les 10 plus proches. La précision est calculée selon deux critères : le voisin est une variante linguistique de la même unité (acronyme, flexion...) ou une unité équivalente (incluant donc les variantes paradigmatiques).

comme valides que les variantes linguistiques d'une même unité (e.g. cm^3 et *cubed centimeters*), alors que la précision 'unité équivalente' accepte les variantes avec les mêmes dimensions (au sens physique, comme cm^3 and *liters*) et suppose donc qu'une conversion est possible et serait nécessaire pour normaliser plus avant. Pour évaluer l'intérêt de cette approche pour normaliser les unités, nous mesurons aussi la précision de l'expression la plus fréquente parmi les 10 ppv (P@1_freq).

Les résultats sont fournis dans le tableau 2. Les bonnes précisions obtenues montrent l'intérêt de cette approche simple pour collecter les variantes. Plus intéressant, ce ne sont pas seulement de simples variantes linguistiques qui sont capturées, mais aussi les variantes paradigmatiques. La très forte P@1_freq en 'unité équivalente' montre le potentiel de cette approche pour aider à normaliser les mesures cliniques détectées avec une unité standard. Voici quelques exemples d'expressions d'unités ainsi regroupées ; celle soulignée est la requête, les autres sont celles retrouvées, par ordre de similarité cosinus décroissante, par notre approche et la plus fréquente est en gras :

— pack of cigarette per day : **cigarette per day** ; cigarette a day ; cigarette daily ; cigarette or equivalent per day ; cigarette/day ; pack per day ; cig/day ; pack/year ; pipe per day ; pack year ; cigarette per week ;

— kgs : pound ; Kg ; lbs ; % of ideal body weight ; lb ; % of ideal weight ; **kg** ; % body weight ; % of their ideal body weight ; % of body weight ; kilogram.

5 Conclusion et perspectives

Nous proposons une approche automatique pour extraire les valeurs numériques et les informations associées (unité, concept, qualificateur, temporalité) à partir des critères d'inclusion des essais cliniques. Les données traitées sont en anglais. Une annotation de référence est obtenue grâce aux annotations effectuées par plusieurs annotateurs. L'accord inter-annotateur va de 0,47 à 0,93, selon les étapes d'annotation et l'expérience des annotateurs. Le système d'extraction exploite ces données de référence et les CRF, et montre une bonne performance : entre 0,65 (concepts) et 0,88 (temps). Notons que le taux d'erreurs diminue avec l'augmentation des données de référence utilisées. Une deuxième étape de notre approche propose de normaliser les unités extraites et montre une précision entre 45 et 100 %. Nous pensons que notre travail peut permettre de traiter et de proposer une représentation des critères d'inclusion, ce qui peut fournir une bonne base pour le recrutement des patients pour les essais cliniques.

Nous avons plusieurs perspectives à ce travail. Nous allons construire un système de conversion des unités extraites en unités standard, comme par exemple *cell/mm3* au lieu de *cell/cm3*, ou *ml* au lieu de *l* ou μl. Concernant les annotations de référence, nous allons préparer un jeu d'annotations plus grand. En effet, nous avons montré qu'il est possible d'obtenir une amélioration des résultats

avec des données d'entraînement plus volumineuses. Le modèle CRF et les règles de conversion seront exploités pour transformer les informations extraites en une représentation formelle des critères d'inclusion et d'exclusion. Selon les travaux de l'état de l'art, il apparaît que ceci correspond actuellement à un défi réel lorsque l'on travaille avec les données en langue naturelle. Finalement, le modèle d'annotation automatique de même que les systèmes de normalisation et de conversion seront appliqués aux essais cliniques et aux dossiers patients pour détecter les patients candidats et les proposer au recrutement. Les hôpitaux en France et au Brésil seront les lieux de test de ce système. Ainsi, nous allons pouvoir appliquer et tester notre approche dans un contexte multilingue. Plus spécifiquement, le modèle d'annotation sera adapté à ces deux langues (français et portugais brésilien), tandis que nous pensons que la présentation des informations numériques respecte les conventions similaires dans différentes langues dans le domaine médical.

Remerciements

Ce travail a été partiellement réalisé et financé dans le cadre du projet Franco-Brésilien CNRS-CONFAP FIGTEM et a bénéficié d'une aide de l'État attribuée au labex COMIN LABS et gérée par l'Agence Nationale de la Recherche au titre du programme « Investissements d'avenir » portant la référence ANR-10-LABX-07-01. Nous tenons également a remercier Guillaume Bouzillé (Univ. Rennes 1), Lucas Oliveira (PUCPR) et Claudia Moro (PUCPR) pour leur aide dans la préparation des données.

Références

ARTSTEIN R. & POESIO M. (2008). Inter-coder agreement for computational linguistics. *Computational Linguistics*, **34**(4).

BIGEARD E., JOUHET V., MOUGIN F., THIESSARD F. & GRABAR N. (2015). Automatic extraction of numerical values from unstructured data in EHRs. In *MIE (Medical Informatics in Europe) 2015*, Madrid, Spain.

CAMPILLO-GIMENEZ B., BUSCAIL C., ZEKRI O., LAGUERRE B., LE PRISÉ E., DE CREVOISIER R. & CUGGIA M. (2015). Improving the pre-screening of eligible patients in order to increase enrollment in cancer clinical trials. *Trials*, **16**(1), 1–15.

CENTER WATCH (2013). *State of the Clinical Trials Industry : A Sourcebook of Charts and Statistics*. Rapport interne, Center Watch.

CONSTANT M., TELLIER I., DUCHIER D., DUPONT Y., SIGOGNE A. & BILLOT S. (2011). Intégrer des connaissances linguistiques dans un CRF : Application à l'apprentissage d'un segmenteur-étiqueteur du français. In *Traitement Automatique du Langage Naturel (TALN'11)*, Montpellier, France.

CUGGIA M., BESANA P. & GLASSPOOL D. (2011). Comparing semi-automatic systems for recruitment of patients to clinical trials. *International Journal of Medical Informatics*, **80**(6), 371–88.

DAVIDOV D. & RAPPAPORT A. (2010). Extraction and approximation of numerical attributes from the web. In *48th Annual Meeting of the Association for Computational Linguistics*, p. 1308–1317.

EMBI P., JAIN A., CLARK J. & HARRIS C. (2005). Development of an electronic health record-based clinical trial alert system to enhance recruitment at the point of care. In *AMIA Symposium*, p. 231–35.

FLETCHER B., GHEORGHE A., MOORE D., WILSON S. & DAMERY S. (2012). Improving the recruitment activity of clinicians in randomised controlled trials : A systematic review. *BMJ Open*, **2**(1), 1–14.

LAFFERTY J., MCCALLUM A. & PEREIRA F. (2001). Conditional random fields : Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning (ICML)*.

LAVERGNE T., CAPPÉ O. & YVON F. (2010). Practical very large scale CRFs. In *Proceedings the 48th Annual Meeting of the Association for Computational Linguistics (ACL)*, p. 504–513 : Association for Computational Linguistics.

MADAAN A., MITTA A., MAUSAM, RAMAKRISHNAN G. & SARAWAGI S. (2016). Numerical relation extraction with minimal supervision. In *Thirtieth AAAI Conference on Artificial Intelligence*.

NATH C., ALBAGHDADI M. & JONNALAGADDA S. (2016). A natural language processing tool for large-scale data extraction from echocardiography reports. *PLoS One*, **11**(4), 153749–64.

OLASOV B. & SIM I. (2006). Ruleed, a web-based semantic network interface for constructing and revising computable eligibility rules. In *AMIA Symposium*, p. 1051.

PRANJAL A., DELIP R. & BALARAMAN R. (2006). Part Of speech Tagging and Chunking with HMM and CRF. In *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest*.

PRESSLER T., YEN P., DING J., LIU J., EMBI P. & PAYNE P. (2012). Computational challenges and human factors influencing the design and use of clinical research participant eligibility pre-screening tools. *BMC Med Inform Dec Mak*, **12**, 47.

RAYMOND C. & FAYOLLE J. (2010). Reconnaissance robuste d'entités nommées sur de la parole transcrite automatiquement. In *Actes de la conférence Traitement Automatique des Langues Naturelles*, Montréal, Canada.

ROSS J., TU S., CARINI S. & SIM I. (2010). Analysis of eligibility criteria complexity in clinical trials. In *Summit on Translational Bioinformatics*, p. 46–50.

SARATH P. R., MANDHAN S. & NIWA Y. (2016). Numerical atrribute extraction from clinical texts. *CoRR*, **1602.00269**.

SCHMID H. (1994). Probabilistic part-of-speech tagging using decision trees. In *Proc of International Conference on New Methods in Language Processing*, p. 44–49.

SEBASTIANI F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, **34**(1), 1–47.

SHIVADE C., RAGHAVAN P., FOSLER-LUSSIER E., EMBI P., ELHADAD N., JOHNSON S. & LAI A. (2014). A review of approaches to identifying patient phenotype cohorts using electronic health records. *J Am Med Inform Assoc*, **21**(2), 221–30.

TU S., PELEG M., CARINI S., BOBAK M., ROSS J., RUBIN D. & SIM I. (2011). A practical method for transforming free-text eligibility criteria into computable criteria. *J Biomed Inform*, **44**(2), 239–50.

WANG T., LI J., DIAO Q., WEI HU Y. Z. & DULONG C. (2006). Semantic event detection using conditional random fields. In *IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW '06)*.